Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

RIDUZIONE DELLA DISTORSIONE IN MODELLI PER LA CLASSIFICAZIONE SCORRETTA DI DATI BINARI

Relatore Prof. Alessandra Salvan
Dipartimento di Scienze Statistiche

Laureanda: Alice Girardi
Matricola N. 1035586

Anno Accademico 2013/2014

Le statistiche sono una forma
di realizzazione del desiderio,
proprio come i sogni.
Jean Baudrillard

Indice

Introduzione

Nell'ambito della modellazione statistica di dati binari, può sorgere il problema della classificazione scorretta. Per errata classificazione si intende l'assegnazione dell'oggetto di studio (un individuo, un valore, un attributo) ad una categoria diversa da quella alla quale dovrebbe essere attribuito. Ignorare l'errata classificazione della variabile risposta può condurre a distorsioni asintotiche non nulle degli stimatori di massima verosimiglianza. Tra i metodi disponibili per correggere la distorsione del primo ordine, sembra particolarmente interessante quello proposto da Firth (1993). Tale metodo consiste in una correzione per l'equazione di verosimiglianza basata sulla funzione di punteggio, piuttosto che per la stima stessa. Firth (1993) mostra che, nei modelli statistici parametrici regolari, il termine dominante della distorsione asintotica dello stimatore di massima verosimiglianza può essere rimosso tramite un'appropriata modificazione della funzione di punteggio. Il vantaggio di tale approccio, oltre la semplice riduzione della distorsione, va identificato nel fornire, in talune situazioni problematiche, una funzione di verosimiglianza con massimo finito.
L'obiettivo principale della tesi consiste nell'illustrare l'efficacia del metodo di Firth (1993) per la riduzione della distorsione in modelli per la classificazione scorretta di dati binari. Il modello di riferimento è quello descritto ad esempio in McInturff et al. (2004), i quali hanno considerato un test dicotomico per diagnosticare un certo stato di malattia D, giungendo alla conclusione che ignorare l'errata classificazione della risposta può condurre a distorsioni asintotiche non nulle degli stimatori di massi-

ma verosimiglianza.
Nel primo capitolo si richiamano definizioni e concetti di base dell'inferenza statistica. Si passa poi ai metodi per la riduzione della distorsione degli stimatori di massima verosimiglianza, al fine di presentare l'approccio di Firth (1993).
Nel secondo capitolo, dopo aver fornito la definizione di errata classificazione per dati binari ed aver presentato le conseguenze sull'inferenza del non tener conto dell'errata classificazione, si passano in rassegna i metodi per la correzione dei conseguenti errori. In particolare, si applica il metodo di Firth (1993) per ridurre la distorsione.
Nel terzo capitolo vengono presentati i risultati degli studi di simulazione effettuati tramite l'ambiente R, relativi alla regressione logistica e alla regressione probit. Mettendo a confronto la distorsione ottenuta a partire dalla funzione di punteggio non modificata con la distorsione ottenuta a partire dalla funzione di punteggio modificata, viene mostrata la validità dell'approccio di Firth (1993) nel ridurre la distorsione dello stimatore di massima verosimiglianza.
Nell'appendice sono stati riportati i comandi R utilizzati per effettuare gli studi di simulazione del Capitolo 3.
Al termine di ogni capitolo si riportano i riferimenti bibliografici relativi al capitolo stesso. La bibliografia complessiva compare alla fine della trattazione.

Capitolo 1

Inferenza di verosimiglianza e metodo di Firth

In questo primo capitolo si richiamano alcune definizioni e concetti di base dell'inferenza basata sulla verosimiglianza, al fine di presentare il metodo proposto da Firth (1993). Tale metodo consiste in una modificazione della funzione di punteggio (*score*) tale da eliminare il termine dominante della distorsione dello stimatore di massima verosimiglianza.

1.1 Modello statistico parametrico

L'inferenza statistica si basa sull'idea che un'indagine o un esperimento empirico abbia generato una collezione di dati, chiamati **campione** ed indicati sinteticamente con y. Nella maggior parte dei casi y è costituito da un vettore di valori, $y = (y_1, y_2, \ldots, y_n)^\top$, ma può anche trattarsi di una struttura più complessa.
Fare inferenza significa considerare y come determinazione di una certa variabile casuale Y ed utilizzare y per trarre conclusioni sulla distribuzione P_0 di Y. La procedura che ha generato y si considera dunque un **esperimento casuale**.
Dal momento che la natura di Y è casuale, le conclusioni sulla sua di-

stribuzione sono soggette ad incertezza. Per questo bisogna fare in modo che:

- il grado di incertezza sia il più piccolo possibile, compatibilmente con la natura casuale di Y;
- sia possibile valutare il grado di incertezza a cui si è sottoposti.

L'insieme delle possibili alternative per la distribuzione di Y è definito dalla natura del fenomeno che ha generato i dati y, dallo schema di campionamento adottato e da altre eventuali informazioni sul fenomeno. Si dice che tale insieme, indicato con $\mathcal{F}$, rappresenta il **modello statistico**. L'inferenza sarà tanto più accurata quanto meglio si saprà delimitare la classe $\mathcal{F}$, compatibilmente con la condizione $P_0 \in \mathcal{F}$.
Il modello statistico $\mathcal{F}$ può essere rappresentato da un qualunque insieme di funzioni di ripartizione, ma esite una situazione che riveste un'importanza cruciale dal punto di vista sia teorico che applicativo. Si tratta del caso in cui gli elementi di $\mathcal{F}$ sono tutte funzioni dello stesso tipo, distinte tra loro solo dal valore θ, libero di variare entro l'insieme $\Theta \subseteq \mathbb{R}^p$ per qualche $p = 1, 2, \ldots$. Allora

$$\mathcal{F} = \{P(\,\cdot\,;\theta), \theta \in \Theta \subseteq \mathbb{R}^p\}$$

dove, per ogni θ fissato, $P(\,\cdot\,;\theta)$ è una funzione di ripartizione su $\mathbb{R}^n$, con p e n numeri naturali.
In molti casi tali funzioni di ripartizione si riferiscono tutte ad una variabile casuale discreta (nel qual caso $\mathcal{F}$ può essere specificata tramite le corrispondenti funzioni di probabilità) o ad una variabile casuale continua (nel qual caso $\mathcal{F}$ può essere specificata tramite le corrispondenti funzioni di densità). Nel seguito della trattazione si utilizzerà l'espressione 'funzione di densità' in entrambi i casi e si scriverà

$$\mathcal{F} = \{p(\,\cdot\,;\theta), \theta \in \Theta \subseteq \mathbb{R}^p\}$$

dove p è una funzione di densità o di probabilità. Il valore θ è detto **parametro**, l'insieme Θ è detto **spazio parametrico** e la classe $\mathcal{F}$ è detta **modello statistico parametrico**.

1.2 Stimatori e stime

Si definisce **statistica** una qualunque funzione del campione casuale che non dipende da altre quantità incognite. Dunque, si può dire che lo **stimatore** è una statistica definita sul campione casuale estratto dalla popolazione di riferimento, mentre la **stima** è il suo corrispondente numerico calcolato sul campione osservato.
Sia $Y \sim p(y;\theta)$ una variabile casuale di cui è nota la famiglia parametrica di appartenenza, ossia di cui si conosce la funzione di densità $p \in \mathcal{F}$, ma non è noto il parametro $\theta \in \Theta$. Da tale variabile casuale si estrae il campione casuale y di dimensione n e sia $\mathcal{Y}$ l'insieme dei valori possibili di y (spazio campionario).
L'obiettivo primario della teoria della stima puntuale consiste nell'individuare un valore di θ che meglio di altri giustifica i dati osservati y. Si ricerca quindi un'opportuna applicazione $\hat{\theta} : \mathcal{Y} \to \Theta$, che fa corrispondere ad ogni $y \in \mathcal{Y}$ un valore $\hat{\theta} = \hat{\theta}(y)$ in Θ. Tale valore è detto **stima** di θ. La statistica $\hat{\theta}(Y)$ è detta, invece, **stimatore** di θ.
La conoscenza della distribuzione campionaria dello stimatore $\hat{\theta}(Y)$, informativa sull'incertezza insita nel processo di stima, è essenziale sia per valutare la bontà di una particolare procedura di stima, sia per confrontare tra loro stimatori alternativi.

1.3 Ordini di successioni

In Matematica è comune l'osservazione che il problema oggetto di studio si semplifica notevolmente quando una variabile o un parametro assumono valori *grandi* o sono prossimi a valori paricolari. Il semplice risultato

che si ottiene può essere usato come risultato di approssimazione per tutte quelle situazioni che non discostano eccessivamente dal caso privilegiato. Talvolta il grado di aderenza della situazione semplificata alla situazione effettiva può essere studiato quantitativamente, tramite opportune disuguaglianze. Più spesso, tuttavia, può essere utile adottare anche uno studio qualitativo.
I simboli $O(\cdot)$ e $o(\cdot)$ costituiscono un'opportuna notazione per gli ordini d'errore per quantità non stocastiche e facilitano lo studio qualitativo dell'errore insito in un risultato di approssimazione. Tale notazione è stata successivamente estesa al caso stocastico da Mann e Wald (1943).

Una successione di reali a_n si definisce **asintoticamente di ordine** $o(n^a)$ se la successione a_n/n^a è infinitesima, ossia se

$$\lim_{n\to+\infty} \frac{a_n}{n^a} = 0.$$

Spesso si omette l'avverbio *asintoticamente*. Se a_n è infinitesima, come ad esempio è la successione n^{-1}, essa è asintoticamente di ordine $o(1)$. La definizione non richiede che la successione abbia limite. Inoltre, essa stabilisce una relazione d'ordine che non va necessariamente pensata come la più stretta possibile: si può dire ugualmente, infatti, che n^{-1} è di ordine $o(n)$ oppure di ordine $o(n^{-1/2})$.
Una successione di reali a_n si definisce **asintoticamente di ordine** $O(n^a)$ se la successione a_n/n^a è limitata, ossia se esiste un reale A tale che, per ogni $n \in \mathbb{N}$,

$$\left|\frac{a_n}{n^a}\right| < A.$$

Tale notazione non richiede né la convergenza di a_n, né la convergenza di a_n/n^a. Se una successione a_n converge ad un limite finito diverso da zero, allora è asintoticamente di ordine $O(1)$, ma non di ordine $o(1)$. Inoltre, una tale successione a_n è anche, per dire, sia di ordine $O(n)$, sia di ordine $o(n)$. Pertanto $O(1)$ è la miglior qualificazione di ordine asintotico per le successioni convergenti non infinitesime. Infine è ovvio che, se a_n è di ordine $o(n^a)$, allora è anche di ordine $O(n^a)$.

1.4 Teoria della verosimiglianza

R.A. Fisher (Londra, 1890 - Adelaide, 1962) è lo statistico, matematico e biologo britannico che ha fatto della statistica una scienza moderna, in quanto fondatore dei concetti di riferimento della statistica matematica moderna. Egli, tra il 1921 ed il 1935, ha introdotto una serie di procedure di inferenza mirate a trattare i modelli statistici parametrici. Tali procedure derivano da un unico elemento di base, molto naturale e di facile comprensione: la funzione di verosimiglianza.

1.4.1 Verosimiglianza e log-verosimiglianza

Sia $\mathcal{F}$ un modello statistico parametrico per i dati y. La funzione $p(y;\theta)$, vista come funzione sia di y che di θ, è detta funzione del modello, con $\theta = (\theta_1, \ldots, \theta_p) \in \Theta \subseteq \mathbb{R}^p$. Fissato il valore y, la funzione del modello $p(y;\theta)$ può essere considerata come funzione solo di θ. Si chiama **funzione di verosimiglianza** (*likelihood function*) di θ basata sui dati y la funzione $L : \Theta \to \mathbb{R}^+$ definita da

$$L(\theta) = p(y;\theta).$$

Talvolta, quando occorre sottolineare la dipendenza della funzione di verosimiglianza dai dati campionari y, tale scrittura viene sostituita da $L(\theta; y)$. Alla luce delle osservazioni, $\theta^1 \in \Theta$ è più credibile di $\theta^2 \in \Theta$ come indice del modello probabilistico generatore dei dati se $L(\theta^1) > L(\theta^2)$. Tramite il rapporto $L(\theta^1)/L(\theta^2)$ il sostegno empirico che $\theta^1 \in \Theta$ riceve da y viene confrontato con quello ricevuto da $\theta^2 \in \Theta$. Due funzioni di verosimiglianza che differiscono solo per una costante moltiplicativa (fattore che non dipende dal parametro θ) sono tra loro equivalenti.
Se si dispone di osservazioni campionarie indipendenti ed identicamente distribuite o di campionamento casuale semplice con numerosità n, allora la funzione di verosimiglianza viene scritta come

$$L(\theta) = \prod_{i=1}^{n} p_i(y_i;\theta),$$

dove p_i è la densità della singola osservazione. Dunque, la verosimiglianza totale si ottiene moltiplicando tra loro le funzioni di verosimiglianza ottenute nei singoli esperimenti.

Dal momento che $L(\theta)$ è una quantità non negativa, e spesso risulta anzi positiva su tutto lo spazio parametrico, le procedure di inferenza statistica basate su $L(\theta)$ sono espresse mediante la **funzione di log-verosimiglianza** (*log-likelihood function*), definita come

$$l(\theta) = \log L(\theta),$$

dove, per convenzione, se $L(\theta) = 0$, si assume che $l(\theta) = -\infty$. Se in luogo della funzione di verosimiglianza si utilizza la sua trasformazione monotona crescente logaritmica, allora l'esecuzione pratica dei calcoli e la derivazione dei risultati teorici risultano più semplici.
I valori del parametro θ caratterizzati da elevata verosimiglianza presentano anche elevata log-verosimiglianza. Così, tramite la differenza $l(\theta^1) - l(\theta^2)$ il sostegno empirico che $\theta^1 \in \Theta$ riceve da y viene confrontato con quello ricevuto da $\theta^2 \in \Theta$. Due funzioni di log-verosimiglianza che differiscono solo per una costante additiva (che non dipende dal parametro θ) sono tra loro equivalenti.
Se si dispone di osservazioni campionarie indipendenti ed identicamente distribuite o di campionamento casuale semplice con numerosità n, allora la funzione di log-verosimiglianza viene scritta come

$$l(\theta) = \sum_{i=1}^{n} \log p_i(y_i; \theta).$$

1.4.2 Stima di massima verosimiglianza

Per una funzione di verosimiglianza $L(\theta)$, con $\theta \in \Theta$, si definisce **stima di massima verosimiglianza** (SMV) di θ un valore $\hat{\theta} \in \Theta$ tale che $L(\hat{\theta}) \geq L(\theta)$ per ogni $\theta \in \Theta$. In altre parole, la stima di massima verosimiglianza di θ è quel valore $\hat{\theta} \in \Theta$ che rende massima la funzione di verosimiglianza $L(\theta)$

sullo spazio parametrico Θ:

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta).$$

Si può determinare $\hat{\theta}$ anche a partire dalla funzione di log-verosimiglianza di θ, di cui costituisce un massimo.
La motivazione intuitiva della stima di massima verosimiglianza è che ''(a parità di altri fattori) scegliamo il sistema [valore del parametro] che dà la massima probabilità ai fatti che abbiamo osservato" (Ramsey, 1931, p. 209).

Occorre, in generale, evidenziare alcune caratteristiche della stima di massima verosimiglianza.

1. Non è strettamente necessario che Θ sia un insieme numerico, cioè che si sia in presenza di un modello parametrico, ma qui verrà considerato solo questo caso.

2. Non è detto che la stima di massima verosimiglianza esista.

3. Se ci sono differenti valori di θ che massimizzano $L(\theta)$, allora la SMV non è unica.

4. La funzione di verosimiglianza va massimizzata sullo spazio parametrico Θ specificato e non per l'intero spazio dei valori di θ che danno un senso matematico a $L(\theta)$.

5. Spesso $\hat{\theta}$ non si può esprimere esplicitamente come funzione dei dati campionari. Quindi, anche se al variare di y in $\mathcal{Y}$ la stima di massima verosimiglianza definisce implicitamente una funzione da $\mathcal{Y}$ in Θ, tale funzione (lo stimatore) non si può rappresentare esplicitamente.

6. Se lo stimatore non consente una rappresentazione esplicita, bisogna ottenere la SMV per via numerica, per il valore osservato di y.

Tuttavia, nella maggior parte dei casi di interesse, la stima di massima verosimiglianza esiste ed è unica. Dato il modello statistico, ogni campione osservabile y dà luogo ad una particolare funzione di verosimiglianza $L(\theta; y)$.
La variabile casuale $\hat{\theta} = \hat{\theta}(Y)$ si definisce **stimatore di massima verosimiglianza** se $\hat{\theta} = \hat{\theta}(y)$ esiste unico con probabilità uno. Dal momento che molte proprietà dello stimatore sono asintotiche, per parlare di stimatore di massima verosimiglianza è sufficiente richiedere che $\hat{\theta}(y)$ esista unico con probabilità che tende ad uno per $n \longrightarrow +\infty$.

1.5 Le famiglie esponenziali

Le famiglie esponenziali costituiscono una classe di modelli statistici parametrici. Si distinguono famiglie esponenziali monoparametriche e famiglie esponenziali multiparametriche.

Una **famiglia esponenziale monoparametrica** è un modello statistico parametrico $\mathcal{F}$ per un'osservazione y, univariata o multivariata, con parametro $\theta \in \Theta \subseteq \mathbb{R}$, e con funzione del modello

$$p(y; \theta) = c(\theta)h(y)\exp\{\psi(\theta)t(y)\},$$

dove $h(\cdot) \geq 0$, $\psi(\theta)$ è una funzione reale di θ con dominio Θ, che non dipende da y, e $t(\cdot)$ è una statistica che non dipende da θ.
Le distribuzioni in $\mathcal{F}$ sono o tutte discrete o tutte assolutamente continue. $c(\theta) \in (0; +\infty)$ è la costante di normalizzazione associata alla funzione $h(y)\exp\{\psi(\theta)t(y)\}$, integrabile e non negativa. Il supporto di Y sotto θ, essendo rappresentato dalla chiusura dell'insieme $\{y \in \mathbb{R}^p : h(y) > 0\}$, è lo stesso per tutte le leggi di probabilità di una data famiglia esponenziale monoparametrica.
Si dice che $\mathcal{F}$ è non banale e che θ è un parametro identificabile se Θ contiene almeno due elementi e $\psi(\cdot)$ è iniettiva. Allora, $\psi = \psi(\theta)$ è un **parametro**

canonico di $\mathcal{F}$ e $t = t(y)$ è una **statistica canonica** di $\mathcal{F}$. In molti casi, Θ è un intervallo in $\mathbb{R}$, eventualmente illimitato, e $\psi(\cdot)$ è derivabile assieme alla sua inversa.
Sono esempi di famiglie esponenziali monoparametriche per un'osservazione y univariata le distribuzioni binomiali con indice m fissato e parametro $\pi \in (0, 1)$, le distribuzioni di Poisson con media $\lambda > 0$, le distribuzioni esponenziali con tasso di guasto $\lambda > 0$, le distribuzioni gamma con parametro di forma fissato e parametro di scala $\lambda > 0$, le distribuzioni gamma con parametro di scala fissato e parametro di forma $\alpha > 0$, le distribuzioni normali univariate con varianza fissata e parametro $\mu \in \mathbb{R}$, le distribuzioni normali univariate con media fissata e parametro $\sigma^2 > 0$.

Una **famiglia esponenziale multiparametrica** è un modello statistico parametrico $\mathcal{F}$ per un'osservazione y, univariata o multivariata, con parametro $\theta \in \Theta \subseteq \mathbb{R}^p$, dove $p > 1$, e con funzione del modello

$$p(y; \theta) = c(\theta)h(y) \exp\{\sum_{j=1}^{k} \psi_j(\theta)t_j(y)\},$$

dove $h(\cdot) \geq 0$ e $\psi(\theta) = (\psi_1(\theta), \ldots, \psi_k(\theta))$ è una funzione di θ con dominio Θ e codominio $\Psi = \psi(\Theta) \subseteq \mathbb{R}^k$.
Le distribuzioni in $\mathcal{F}$ sono o tutte discrete o tutte assolutamente continue. $c(\theta) \in (0; +\infty)$ è la costante di normalizzazione associata alla funzione $h(y) \exp\{\sum_{j=1}^{k} \psi_j(\theta)t_j(y_i)\}$, integrabile e non negativa. Il supporto di Y sotto θ, essendo rappresentato dalla chiusura dell'insieme $\{y \in \mathbb{R}^p : h(y) > 0\}$, è lo stesso per tutte le leggi di probabilità di una data famiglia esponenziale multiparametrica.
Si dice che $\theta = (\theta_1, \ldots, \theta_p)$ è un parametro identificabile se $\psi(\theta)$ è iniettiva. $p(y; \theta) = c(\theta)h(y) \exp\{\sum_{j=1}^{k} \psi_j(\theta)t_j(y)\}$ è una **rappresentazione minimale**, ossia coinvolge il minimo numero di funzioni $\psi_j(\theta)$ e di associate statistiche $t_j(y)$, se sono soddisfatte le seguenti tre condizioni:

1. Θ deve contenere almeno $k + 1$ elementi;

2. le $k+1$ funzioni reali $1, \psi_1(\theta), \ldots, \psi_k(\theta)$ devono essere linearmente indipendenti, cioè, per ogni $\theta \in \Theta$, deve essere $\psi_k(\theta) = c_0 + c_1\psi_1(\theta) + \ldots + c_{k-1}\psi_{k-1}(\theta)$ in modo da poter riscrivere la funzione del modello utilizzando solo $\psi_j(\theta)$ con le statistiche associate $t_j'(y) = t_j(y) + c_j t_k(y)$ per $j = 1, \ldots, k-1$;

3. le $k+1$ funzioni reali $1, t_1(y), \ldots, t_k(y)$ devono essere linearmente indipendenti.

Quando una famiglia esponenziale è scritta in una rappresentazione minimale, k è l'**ordine** della famiglia e $t = t(y) = (t_1(y), \ldots, t_k(y))$ è una **statistica canonica** di $\mathcal{F}$.
Si dice che $\psi = \psi(\theta)$ è un **parametro canonico** se l'ordine della famiglia coincide con la dimensione di Θ, cioè $k = p$, e $\psi(\theta)$ è una riparametrizzazione del modello, con $\psi(\cdot)$ differenziabile su intΘ, come pure l'inversa $\theta = \theta(\psi)$ su intΨ. Nella parametrizzazione canonica ψ, la statistica canonica t ha densità del tipo

$$p(t;\psi) = c(\theta(\psi))\tilde{h}(t)\exp\{\sum_{j=1}^{p}\psi_j t_j\},$$

con $\psi \in \Psi$ e $\tilde{h}(t)$ opportuna. La famiglia esponenziale con densità $p(t;\psi)$ è definita **regolare** se Ψ è un insieme aperto e contiene tutti i valori $\psi \in \mathbb{R}^p$ per cui la funzione non negativa $\tilde{h}(t)\exp\{\sum_{j=1}^{p}\psi_j t_j\}$ risulta integrabile.
Sono esempi di famiglie esponenziali multiparametriche di ordine 2 per un'osservazione y univariata le distribuzioni normali univariate con parametro $\theta = (\mu, \sigma^2)$, dove $\mu \in \mathbb{R}$ e $\sigma^2 > 0$, e le distribuzioni gamma con parametro $\theta = (\alpha, \lambda)$, dove $\alpha > 0$ e $\lambda > 0$. Sono esempi di famiglie esponenziali multiparametriche di ordine maggiore di 2 le distribuzioni normali multivariate e le distribuzioni multinomiali.

1.6 Metodi tradizionali per la riduzione della distorsione

Se si dispone di un modello regolare con parametro $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ p-dimensionale, la distorsione asintotica dello stimatore di massima verosimiglianza $\hat{\theta} = \hat{\theta}(Y)$, basato su un campione di numerosità n, può essere scritta come

$$b(\theta) = E_\theta(\hat{\theta}) - \theta = \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + \ldots.$$

Sono stati ampiamente studiati e discussi due approcci tradizionali (Cox e Hinkley, 1974, § 8.4) per la riduzione della distorsione. Un aspetto comune di questi due metodi è il fatto di essere 'correttivi' piuttosto che 'preventivi', cioè la stima di massima verosimiglianza $\hat{\theta}$ viene prima calcolata e poi corretta.

Il primo metodo (Quenouille, 1949, 1956) non richiede dettagliati calcoli numerici e risulta dunque particolarmente adatto a risolvere problemi complessi. Esso è chiamato *jackknife* (letteralmente **coltello a serramanico**) oppure *sample-splitting* e si basa sulla seguente idea. Sia y un campione casuale semplice con numerosità $n > 1$. Si supponga di avere a disposizione tutte le osservazioni tranne la j-esima. Con queste $n-1$ osservazioni si può ancora stimare il valore del parametro della funzione. Tipicamente tale stima è leggermente diversa da quella ottenuta utilizzando tutte le n osservazioni e la differenza tra queste due stime fornisce proprio l'informazione con cui calcolare l'indeterminazione sulla stima del parametro. Dunque il metodo *jackknife* consiste nel ricalcolare più volte la grandezza statistica stimata lasciando fuori dal campione un'osservazione alla volta. Sia $\hat{\theta}_n$ una stima calcolata a partire da $Y_1, \ldots, Y_n$ e sia $\hat{\theta}_{n-1,j}$ la stessa stima calcolata a partire dall'insieme di $n-1$ variabili casuali ottenuto omettendo Y_j. Sia $\bar{\theta}_{n-1,\cdot}$ la media di $\hat{\theta}_{n-1,j}$ con $j = 1, \ldots, n$. Rispettivamente per $n-1$ e per n si ha

$$E_\theta(\hat{\theta}_{n-1}) = \theta + \frac{b_1(\theta)}{(n-1)} + \frac{b_2(\theta)}{(n-1)^2} + O((n-1)^{-3});$$

$$E_\theta(\bar{\theta}_{n-1,\cdot}) = \theta + \frac{b_1(\theta)}{(n-1)} + O(n^{-2});$$
$$E_\theta(\hat{\theta}_n) = \theta + \frac{b_1(\theta)}{n} + O(n^{-2}).$$

Dalle ultime due equazioni si può ottenere una combinazione lineare di $\bar{\theta}_{n-1,\cdot}$ e $\hat{\theta}_n$ con distorsione di ordine n^{-2}. Infatti, per

$$\hat{\theta}_n^J = n\hat{\theta}_n - (n-1)\bar{\theta}_{n-1,\cdot} = n\hat{\theta}_n - \frac{(n-1)}{n}\sum_{j=1}^{n}\hat{\theta}_{n-1,j}$$

risulta $E_\theta(\hat{\theta}_n^J) = \theta + O(n^{-2})$.
Se $\hat{\theta}_n$ è una media campionaria, allora $\hat{\theta}_n = \bar{\theta}_{n-1,\cdot} = \hat{\theta}_n^J$.
Il metodo *jackknife* richiede solo che il termine principale nella distorsione sia di ordine n^{-1}. Se il termine principale nello sviluppo di $E_\theta(\hat{\theta}_n)$ è di ordine n^{-2}, le modificazioni sopra descritte non sono necessarie; naturalmente, la distorsione di ordine uno non può essere rimossa.
Come già detto, il metodo *jackknife* è principalmente utilizzato nei problemi relativamente complicati come l'analisi dei dati di sopravvivenza, l'analisi delle serie temporali e l'analisi multivariata, in cui un diretto lavoro analitico non è possibile. Ciò nonostante, è utile osservare i risultati che si ottengono applicando il metodo a casi relativamente semplici. Un esempio elementare riguarda la stima della varianza a partire da

$$\hat{\theta}_n = \frac{\sum_{j=1}^{n}(Y_j - \bar{Y}_\cdot)^2}{n},$$

dove $\bar{Y}_\cdot$ indica la media campionaria. Si dimostra che

$$\hat{\theta}_n^J = \frac{\sum_{j=1}^{n}(Y_j - \bar{Y}_\cdot)^2}{(n-1)}.$$

Il secondo approccio tradizionale è chiamato **riduzione della distorsione tramite sviluppo in serie** (*reduction of bias by series expansion*).
Sia $\hat{\theta}(Y)$ lo stimatore di massima verosimiglianza ottenuto da un campione casuale semplice con numerosità n. La media dello stimatore è

$$E_\theta[\hat{\theta}(Y)] = \theta + b(\theta) + O(n^{-2}),$$

dove $b(\theta)$ è di ordine $O(n^{-1})$.
Si può mostrare (Pace e Salvan, 1996, § 9.4.2) che ad esempio nel caso univariato, cioè per $p = 1$,

$$E_\theta(\hat{\theta} - \theta) = \frac{1}{2} i(\theta)^{-2}(\nu_3(\theta) + 2\nu_{2,1}(\theta)) + O(n^{-2}) = b(\theta) + O(n^{-2}),$$

dove $i(\theta) = -E_\theta(\frac{\partial^2 l(\theta)}{\partial \theta^2})$, $\nu_3(\theta) = E_\theta(\frac{\partial^3 l(\theta)}{\partial \theta^3})$ e $\nu_{2,1}(\theta) = E_\theta(\frac{\partial^2 l(\theta)}{\partial \theta^2}\frac{\partial l\theta}{\partial \theta})$.
Semplicemente sostituendo $\hat{\theta}$ al posto del parametro ignoto θ in $\frac{b_1(\theta)}{n}$, la stima corretta per la distorsione del primo ordine risulta quindi essere

$$\hat{\theta}_I = \hat{\theta} - b(\hat{\theta}).$$

Si dimostra che essa ha distorsione $O(n^{-2})$: $E_\theta(\hat{\theta}_I(Y) - \theta) = O(n^{-2})$.

Entrambi questi metodi hanno successo nella rimozione del termine $\frac{b_1(\theta)}{n}$ dalla distorsione asintotica; il primo ha il vantaggio di non richiedere calcoli teorici (anche se ciò è solitamente controbilanciato da una perdita di precisione), mentre lo stimatore $\hat{\theta}_I$ è in generale efficiente (Firth, 1993).
Un requisito fondamentale per l'applicazione dell'uno o dell'altro metodo ad un campione di osservazioni è l'esistenza di $\hat{\theta}$ finito per tale campione (nel caso del *jackknife* $\hat{\theta}$ deve esistere finito anche per tutti i sotto-campioni del campione originario). Nel caso in cui $\hat{\theta}$ sia infinito, come può avvenire ad esempio nei modelli di regressione logistica, i due approcci tradizionali per la riduzione della distorsione non sono applicabili.

Un ulteriore metodo tradizionalmente utilizzato per ridurre la distorsione dello stimatore di massima verosimiglianza è il *bootstrap*, introdotto da Efron (1979). Tale metodo consente di correggere la distorsione quando la forma di $b(\theta)$ non è nota. Partendo da un campione di stime $\hat{\theta}_1, \ldots, \hat{\theta}_R$, si può calcolare la loro media

$$\hat{E}_\theta(\hat{\theta}) = \frac{1}{R}\sum_{i=1}^{R} \hat{\theta}_i.$$

Quando $R \to \infty$,

$$\hat{E}_\theta(\hat{\theta}) \longrightarrow E_{\hat{\theta}}(\hat{\theta}) = \hat{\theta} + b(\hat{\theta}),$$

così che, per R sufficientemente grande, $\hat{E}_\theta(\hat{\theta}) - \hat{\theta}$ è una buona approssimazione di $b(\hat{\theta})$.

1.7 L'approccio di Firth per la riduzione della distorsione

Firth (1993) mostra che nei modelli parametrici regolari, il termine dominante della distorsione asintotica dello stimatore di massima verosimiglianza può essere rimosso anche tramite un'appropriata modificazione della funzione di punteggio (*score function*).
Il metodo proposto da Firth (1993) non è vincolato alla finitezza di $\hat{\theta}$. Esso consiste in una correzione sistematica del meccanismo che produce la stima di massima verosimiglianza, cioè dell'equazione di verosimiglianza basata sulla funzione di punteggio, piuttosto che della stima stessa.

In modelli regolari, la stima di massima verosimiglianza $\hat{\theta}$ del parametro θ si ottiene come soluzione dell'equazione di verosimiglianza:

$$\nabla l(\theta) = U(\theta) = 0.$$

Per ridurre la distorsione asintotica dello stimatore di massima verosimiglianza si effettua una modificazione della funzione di punteggio.
La funzione di punteggio modificata è

$$U^*(\theta) = U(\theta) + A(\theta) = [U_r + A_r],$$

dove $[a_r]$ indica il vettore con generica componente a_r, per $r = 1, \ldots, p$.
La stima di massima verosimiglianza corretta θ^* si ottiene dunque come soluzione dell'equazione di stima basata sulla funzione di punteggio modificata

$$U^*(\theta) = 0,$$

con $A(\theta)$ tale che la distorsione asintotica dello stimatore risulti di ordine $O(n^{-2})$, inferiore rispetto alla distorsione di $\hat{\theta}$.

Si consideri dapprima un modello di tipo esponenziale con funzione di log-verosimiglianza $l(\theta) = t\theta - K(\theta)$ in cui $p = 1$. Si ottiene

$$U(\theta) = l^{'}(\theta) = t - K^{'}(\theta).$$

In questo caso, la statistica sufficiente t non influenza la forma di $U(\theta)$, ma solo la sua posizione. La distorsione di $\hat{\theta}$ deriva dalla combinazione di due fattori:

1. non distorsione della funzione di punteggio, $E_\theta\{U(\theta)\} = 0$ al vero valore di θ;

2. curvatura della funzione di punteggio, $l^{''}(\theta) \neq 0$.

Se $U(\theta)$ è lineare in θ, allora $E_\theta(\hat{\theta}) = \theta$, ma la curvatura e la non distorsione della funzione di punteggio si combinano provocando una distorsione nello stimatore di massima verosimiglianza $\hat{\theta}$.
Dunque la distorsione di $\hat{\theta}$ può essere ridotta attraverso l'introduzione di una piccola distorsione nella funzione di punteggio. La modificazione appropriata per $U(\theta)$ è illustrata nella Fig.1.1. Se $\hat{\theta}$ è soggetto ad una distorsione positiva (cioè se $E_\theta(\hat{\theta}) > \theta$), la funzione di punteggio in media deve essere spostata verso il basso in ogni punto θ di una quantità pari a $i(\theta)b(\theta)$, dove $b(\theta)$ indica la distorsione e $-i(\theta) = E_\theta(U^{'}(\theta))$ rappresenta il gradiente medio, con $U^{'}(\theta) = \frac{\partial U(\theta)}{\partial \theta}$. Si ha infatti

$$U(\theta^* + b(\theta^*)) \doteq U(\theta^*) + b(\theta^*)(-i(\theta^*)),$$

e dunque la traslazione verso il basso di $U(\theta)$ deve essere pari a $A(\theta) = -i(\theta)b(\theta)$.
Si ottiene quindi una nuova funzione di punteggio

$$U^*(\theta) = U(\theta) - i(\theta)b(\theta).$$

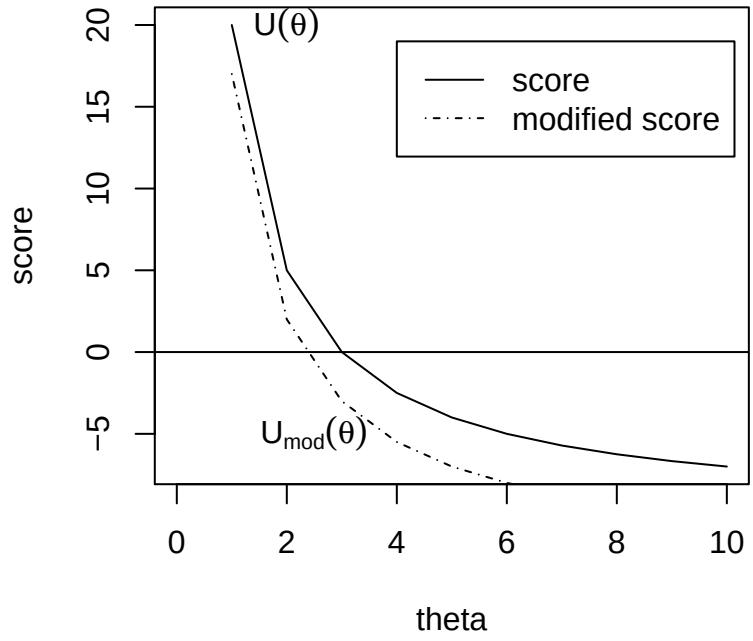

Figura 1.1: *Modificazione della funzione di punteggio, dove* $U_{mod}(\theta)$ *rappresenta* $U^*(\theta)$.

Essa può essere interpretata come un vettore di equazioni nel caso si disponga di un vettore di parametri e $i(\theta)$ è la matrice di informazione attesa (o informazione di Fisher).
Ponendo infine $U^*(\theta) = 0$, si ottiene la stima modificata θ^*.

Per formalizzare l'argomento fin qui discusso ed estenderlo anche ai problemi che non riguardano le famiglie esponenziali si utilizza una diversa notazione per le derivate della funzione di log-verosimiglianza e per i loro momenti nulli (McCullagh, 1987).
Le derivate della funzione di log-verosimiglianza sono indicate con

$$U_r(\theta) = \frac{\partial l(\theta)}{\partial \theta_r}, \qquad U_{rs}(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s},$$

e così via, dove $\theta = (\theta_1, \ldots, \theta_p)$ è il vettore dei parametri.
I momenti nulli sono definiti come

$$\kappa_r = n^{-1} E_\theta\{U_r\}, \qquad \kappa_{rs} = n^{-1} E_\theta\{U_{rs}\},$$

$$\kappa_{rst} = n^{-1} E_\theta\{U_{rst}\}, \qquad \kappa_{rstu} = n^{-1} E_\theta\{U_{rstu}\},$$

$$\kappa_{r,s} = n^{-1} E_\theta\{U_r U_s\}, \qquad \kappa_{r,s,t} = n^{-1} E_\theta\{U_r U_s U_t\},$$

$$\kappa_{r,s,t,u} = n^{-1} E_\theta\{U_r U_s U_t U_u\},$$

$$\kappa_{r,st} = n^{-1} E_\theta\{U_r U_{st}\}, \qquad \kappa_{r,stu} = n^{-1} E_\theta\{U_r U_{stu}\},$$

$$\kappa_{rs,tu} = n^{-1} E_\theta\{U_{rs} U_{tu}\}, \qquad \kappa_{r,s,tu} = n^{-1} E_\theta\{U_r U_s U_{tu}\},$$

e così via. Sono note le seguenti relazioni di Bartlett:

$$\kappa_r = 0,$$

$$\kappa_{rs} + \kappa_{r,s} = 0,$$

$$\kappa_{rst} + \kappa_{r,st} + \kappa_{s,rt} + \kappa_{t,rs} + \kappa_{r,s,t} = 0,$$

$$\kappa_{rst} + \kappa_{r,st}[3] + \kappa_{r,s,t} = 0,$$

$$\kappa_{rstu} + \kappa_{r,stu}[4] + \kappa_{rs,tu}[3] + \kappa_{r,s,tu}[6] + \kappa_{r,s,t,u} = 0.$$

Una funzione di punteggio modificata abbastanza generale è data da

$$U_r^*(\theta) = U_r(\theta) + A_r(\theta)$$

in cui il pedice r indica che la *score function* modificata segue le componenti di θ e A_r può dipendere dai dati. A_r è $O_p(1)$ quando $n \to \infty$, ossia è limitata in probabilità (cf. Pace e Salvan, 1996, §3.4.3). Siano $\hat{\theta}$ e θ^* tali da soddisfare $U(\hat{\theta}) = 0$ e $U(\theta^*) = 0$ e sia $\hat{\gamma} = n^{\frac{1}{2}}(\theta^* - \theta)$. Allora la distorsione di θ^*, basata su uno sviluppo di $U_r^*(\theta^*)$, è

$$E_\theta(n^{-\frac{1}{2}}\hat{\gamma}^r) = n^{-1}\kappa^{r,s}\{-\kappa^{t,u}\frac{(\kappa_{s,t,u} + \kappa_{s,tu})}{2} + \alpha_s\} + O(n^{-\frac{3}{2}}),$$

dove $[\kappa^{r,s}]$ è l'inversa della matrice di informazione attesa $[\kappa_{r,s}]$ e α_s è il valore atteso nullo di A_s.
La distorsione di primo ordine di $\hat{\theta}$ è rappresentata (Firth, 1993) dal termine

$$-n^{-1}\kappa^{r,s}\kappa^{t,u}\frac{(\kappa_{s,t,u} + \kappa_{s,tu})}{2} = n^{-1}b_1^r(\theta).$$

Quindi A_r rimuove il termine di primo ordine se soddisfa la seguente condizione:

$$\kappa^{r,s}\alpha_s = -b_1^r + O(n^{-\frac{1}{2}})$$

da cui si ottiene

$$\alpha_r = -\kappa_{r,s}b_1^s + O(n^{-\frac{1}{2}}).$$

In notazione matriciale, il vettore A dovrebbe essere tale che

$$E_\theta(A) = -i(\theta)\frac{b_1(\theta)}{n} + O(n^{-\frac{1}{2}}).$$

Usando rispettivamente l'informazione attesa e l'informazione osservata, ovvii candidati per una scelta di A che riduce la distorsione sono dunque $A^{(E)} = -i(\theta)\frac{b_1(\theta)}{n}$ e $A^{(O)} = -I(\theta)\frac{b_1(\theta)}{n}$, con $I(\theta) = -U'(\theta)$. Nel caso di famiglia esponenziale con parametrizzazione canonica $A^{(O)}$ e $A^{(E)}$ coincidono in quanto l'informazione osservata non dipende dai dati.
In generale, entrambe le modificazioni $A^{(E)}$ e $A^{(O)}$ rimuovono il termine di distorsione di ordine $O(n^{-1})$.

Il vantaggio del metodo proposto da Firth (1993) sembra andare oltre la riduzione della distorsione e va individuato soprattutto nel fornire, in talune situazioni problematiche, una funzione di verosimiglianza con massimo finito.

1.8 Applicazione del metodo alle famiglie esponenziali

1.8.1 La distribuzione a priori di Jeffreys come funzione di penalità per la riduzione della distorsione

Se θ è il parametro canonico di un modello di una famiglia esponenziale, $\kappa_{r,st} = 0$ per ogni r, s e t. Dunque, utilizzando la **convenzione per la somma** di Einstein (Pace e Salvan, 1996, §9.4.3), in base alla quale, quando in un prodotto di elementi di matrici generalizzate un indice compare due o più volte, si sottintende la somma rispetto a quell'indice sopra il campo di variazione (sottinteso perché ovvio), l'elemento r-esimo di $A^{(E)}(\theta)$ (o equivalentemente di $A^{(O)}(\theta)$) è dato da

$$A_r = -n\kappa_{r,s}\frac{b_1^s}{n} = \frac{\kappa_{r,s}\kappa^{s,t}\kappa^{u,v}\kappa_{t,u,v}}{2} = \frac{\kappa^{u,v}\kappa_{r,u,v}}{2} = \frac{-\kappa^{u,v}\kappa_{ruv}}{2}$$

che, in notazione matriciale, diventa

$$A_r = \frac{1}{2}tr\{i^{-1}(\frac{\partial i}{\partial\theta_r})\} = \frac{\partial}{\partial\theta_r}\{\frac{1}{2}\log|i(\theta)|\}.$$

La soluzione di $U_r^* \equiv U_r + A_r = 0$, individua dunque un punto stazionario di

$$l^*(\theta) = l(\theta) + \frac{1}{2}\log|i(\theta)|$$

o, equivalentemente, della funzione di verosimiglianza penalizzata

$$L^*(\theta) = L(\theta)|i(\theta)|^{\frac{1}{2}}.$$

Il determinante dell'informazione di Fisher elevato alla $\frac{1}{2}$ si chiama **distribuzione a priori di Jeffreys**. Per il parametro canonico di un modello di una famiglia esponenziale il termine di distorsione $O(n^{-1})$ è dunque rimosso dallo stimatore definito come moda della distribuzione a posteriori basata su questa distribuzione a priori.

1.9 Espressione generale

Si consideri ora uno scenario più generale che include sia i modelli della famiglia esponenziale con parametrizzazione non canonica, sia i modelli non esponenziali.
La funzione di punteggio modificata può essere scritta come

$$U_r^* = U_r + A_r,$$

dove $A_r(\theta)$ è basato o sull'informazione attesa, nel qual caso si ha

$$A_r = A_r^{(E)} = n\kappa_{r,s}\kappa^{s,t}\kappa^{u,v}\frac{(\kappa_{t,u,v} + \kappa_{t,uv})}{2n} = \kappa^{u,v}\frac{(\kappa_{r,u,v} + \kappa_{r,uv})}{2},$$

o sull'informazione osservata, per cui

$$A_r = A_r^{(O)} = -U_{rs}\kappa^{s,t}\kappa^{u,v}\frac{(\kappa_{t,u,v} + \kappa_{t,uv})}{2n}.$$

Le stime derivate utilizzando $A_r^{(O)}$ risultano preferibili in termini di efficienza. Per dimostrarlo, Firth (1993) ha considerato uno sviluppo di $U^*(\theta^*)$. In particolare, dalla definizione,

$$0 = U_r^*(\theta^*) = U_r(\theta^*) + A_r(\theta^*).$$

Se $A_r(\theta) = A_r^{(O)}(\theta) = U_{rs}(\theta)\frac{b_1^s(\theta)}{n}$, si ha

$$(\theta^* - \hat{\theta})^r = -\frac{b_1^r(\hat{\theta})}{n} + O_p(n^{-2}),$$

mentre se $A_r(\theta) = A_r^{(E)}(\theta) = -i_{rs}(\theta)\frac{b_1^s(\theta)}{n}$, allora

$$(\theta^* - \hat{\theta})^r = -\frac{b_1^r(\hat{\theta})}{n} - i^{rs}(\hat{\theta})\{U_{st}(\hat{\theta}) + i_{st}(\hat{\theta})\}\frac{b_1^t(\hat{\theta})}{n} + O_p(n^{-2}).$$

La differenza $U_{st}(\hat{\theta}) + i_{st}(\hat{\theta})$ tra l'informazione attesa e l'informazione osservata nella stima di massima verosimiglianza è, in generale, $O_p(n^{-\frac{1}{2}})$, così che il termine supplementare in $(\theta^* - \hat{\theta})^r$ è $O_p(n^{-\frac{3}{2}})$.
Da $(\theta^* - \hat{\theta})^r = -\frac{b_1^r(\hat{\theta})}{n} + O_p(n^{-2})$ si può concludere (Firth, 1993) che, se U^* è calcolato utilizzando l'informazione osservata, θ^* coincide con $\hat{\theta}_I$ al secondo ordine, mentre non è così se si utilizza l'informazione attesa. In generale, l'uso della modificazione $A^{(E)}$ implica una perdita di precisione al secondo ordine rispetto all'uso di $A^{(O)}$.

1.10 Sintesi

Come mostrato, nei problemi regolari, la distorsione asintotica dello stimatore di massima verosimiglianza può essere ridotta mediante la rimozione del termine $O(n^{-1})$, ottenuta introducendo un appropriato termine di distorsione nella funzione di punteggio. Se il parametro di interesse è il parametro canonico di una famiglia esponenziale, ciò equivale ad utilizzare la distribuzione a priori di Jeffreys come funzione di penalità per la verosimiglianza. Nel caso di altre parametrizzazioni, sono disponibili diversi tipi di correzioni, ottenute utilizzando l'informazione attesa o l'informazione osservata. Al di fuori dei modelli della famiglia esponenziale, l'uso dell'informazione attesa si traduce in una perdita di efficienza rispetto all'uso dell'informazione osservata.

1.11 Riferimenti bibliografici

Le parti relative al modello statistico parametrico, definito nel paragrafo 1.1, alla differenza tra stima e stimatore di un parametro, discussa nel paragrafo 1.2, ed alla stima di massima verosimiglianza, presentata nel sottoparagrafo 1.4.2, sono state tratte da Azzalini (2001, §§2.1.1, 2.1.2, 3.1.2).
Il paragrafo 1.2 è stato in parte tratto da Piccolo (2006, §13.2).
Gli ordini di successioni, definiti nel paragrafo 1.3, la riduzione della di-

storsione tramite sviluppo in serie, introdotta nel paragrafo 1.6, e la convenzione per la somma di Einstein, utilizzata nel paragrafo 1.8, sono stati tratti da Pace e Salvan (1996, §§3.4.1, 4.2, 9.4.2, 9.4.3).
I paragrafi 1.4 e 1.5, dedicati rispettivamente alla teoria della verosimiglianza e alle famiglie esponenziali, sono stati tratti da Pace e Salvan (2001, §§3.1, 3.2, A.7.4, A.7.5).
Le sezioni relative alla riduzione della distorsione dello stimatore di massima verosimiglianza sono state tratte da Firth (1993), e si è fatto riferimento anche a Quenouille (1949, 1956), Cox & Hinkley (1974, §8.4), McCullagh (1987) e Scholz (2007).

Capitolo 2

Modelli per la classificazione scorretta di dati binari

In questo capitolo, dopo aver fornito la definizione di errata classificazione per dati binari ed aver presentato le conseguenze sull'inferenza del non tener conto dell'errata classificazione, si passano in rassegna i metodi per la correzione dei conseguenti errori ed il collegato problema delle variabili esplicative binarie classificate scorrettamente. In particolare, si passa all'illustrazione del modello di McInturff et al. (2004), i quali hanno considerato un test per diagnosticare un certo stato di malattia D, giungendo alla conclusione che ignorare l'errata classificazione della risposta può condurre a distorsioni asintotiche non nulle degli stimatori di massima verosimiglianza. Infine, per correggere tali distorsioni si utilizza il metodo di Firth (1993), presentato nel capitolo precedente.

2.1 Una definizione di errata classificazione

Per errata classificazione si intende l'assegnazione dei soggetti in esame alla categoria sbagliata di una variabile categoriale.
La classificazione scorretta è, dunque, un *errore di classificazione*; essa può essere definita come la classificazione di un individuo, un valore, un at-

tributo in una categoria diversa da quella alla quale dovrebbe essere attribuito. In una tabella di contingenza, l'errata classificazione è il passaggio erroneo da una cella ad un'altra.
Il concetto di classificazione scorretta interessa, pertanto, le variabili discrete, non necessariamente binarie, mentre per gli errori che interessano le variabili continue si usa più frequentemente applicare il termine *errore di misura*. In epidemiologia l'errata classificazione può riguardare la malattia, l'esposizione o il fattore di confondimento. Essa può, inoltre, essere di due tipi: differenziale e non differenziale. La classificazione scorretta differenziale è influenzata da altre variabili in esame, mentre la classificazione scorretta non differenziale non dipende da altre variabili in esame.

2.2 Breve rassegna

Parecchi autori hanno esaminato gli effetti di risposte binarie erroneamente classificate (Barron, 1977; Copeland et al., 1977; Fleiss, 1981, Ch.11), i metodi per la correzione dei conseguenti errori (Green, 1983; Greenland, 1988; Franco, 1992; Brenner & Gefeller, 1993; Carroll, Ruppert & Stefanski, 1995, Ch.13) ed il collegato problema delle variabili esplicative binarie classificate scorrettamente (Flegal, Brownie & Haas, 1986; Weinberg, Umbach & Greenland, 1994). Tuttavia, essi si sono focalizzati su una sola variabile esplicativa binaria e su un solo tipo di modello di regressione.
Magder & Hughes (1997) hanno proposto un approccio per incorporare la probabilità di classificazione scorretta nella stima dei coefficienti di regressione. Tale approccio genera stimatori consistenti degli effetti della covariata e degli errori standard associati per ogni modello di regressione binaria.
Neuhaus (1999) ha poi proposto un approccio computazionalmente più efficiente per trattare le risposte misurate con errore. Egli, estendendo ed unificando lavori precedenti, ha derivato espressioni generali per la distorsione che si possono applicare ad ogni tipo di modello di regressione

binaria e ad ogni tipo di variabile esplicativa. Tali espressioni vanno bene sia nel caso in cui la distorsione dovuta alla classificazione scorretta della risposta sia piccola ed ignorabile, sia nel caso in cui quest'ultima sia grande.
Ignorare l'errata classificazione della risposta può condurre a distorsioni non nulle.

2.3 I modelli lineari generalizzati

La classe dei **modelli lineari generalizzati** (GLM), pur non essendo enormemente ampia da un punto di vista prettamente matematico, è tuttavia sufficientemente flessibile da incorporare un gran numero di situazioni importanti per le applicazioni pratiche.
La classe dei modelli lineari generalizzati è stata introdotta da Nelder & Wedderburn (1972), i quali hanno evidenziato il fatto che molti modelli e metodi specifici già in uso facevano in realtà parte di un'unica tipologia.
Dal momento che si considera una categoria di distribuzioni di probabilità che è un'estensione delle famiglie esponenziali, conviene introdurre una notazione specifica per questo genere di distribuzioni. Dunque, per una variabile casuale reale Y si scriverà che

$$Y \sim EF(b(\theta), \psi/\omega)),$$

se Y ha funzione di densità del tipo

$$p(y;\theta,\psi) = \exp\left(\frac{\omega}{\psi}\left\{y\theta - b(\theta)\right\} + c(y,\psi)\right),$$

dove θ e ψ sono parametri scalari ignoti, ω è una costante nota, e $b(\cdot)$ e $c(\cdot)$ sono funzioni note la cui scelta individua una particolare distribuzione di probabilità.
In un modello lineare normale, per la generica unità i-esima, si costruisce un predittore lineare $\eta_i = x_i^\top \beta$ per $i = 1, \ldots, n$. La corrispondente osservazione y_i è tratta da $Y_i \sim N(\mu_i, \sigma^2)$, dove la relazione tra il valore medio μ_i

ed il predittore lineare η_i è di identità. In formule,

$$Y_i \sim N(\mu_i, \sigma^2), \qquad \mu_i = \eta_i, \qquad \eta_i = x_i^\top \beta.$$

La classe dei GLM si ottiene estendendo la formulazione precedente come segue:

- considerando come possibile distribuzione per Y_i non solo la normale, ma una qualsiasi altra distribuzione $EF(b(\theta_i), \psi/\omega_i)$ tale che $b^{'}(\theta_i) = \mu_i$;

- considerando altre forme di legame tra il valore medio μ_i ed il predittore lineare η_i, cioè ipotizzando $g(\mu_i) = \eta_i$ con $g(\cdot)$ funzione monotona derivabile detta funzione di legame (o semplicemente legame) tra μ ed η.

In formule,

$$Y_i \sim EF(b(\theta_i), \psi/\omega_i), \qquad g(\mu_i) = \eta_i, \qquad \eta_i = x_i^\top \beta.$$

Quando la funzione di legame $g(\cdot)$ è l'identità e la struttura dell'errore è di tipo normale si torna ad un modello lineare tradizionale.
Nei modelli lineari generalizzati non si ha più la precisa separazione della variabile risposta nelle due componenti sistematica ed accidentale. Ad ogni modo, per sottolineare la connessione dei GLM con i modelli lineari, anche in questo caso si continua a parlare di termine d'errore, che in realtà non esiste.

2.4 Forma generale del modello

McInturff et al. (2004) hanno considerato un test dicotomico per diagnosticare un certo stato di malattia D. La variabile Z rappresenta lo stato vero, vale a dire, $Z = 1$ indica che la malattia è presente e $Z = 0$ indica che D è

assente. La variabile Y rappresenta il risultato del test, dove $Y = 1$ indica che il test è positivo e $Y = 0$ indica che il test è negativo.

$$Z = \begin{cases} 1 & \text{se la malattia è presente} \\ 0 & \text{se la malattia è assente} \end{cases}$$

$$Y = \begin{cases} 1 & \text{se il test è positivo} \\ 0 & \text{se il test è negativo} \end{cases}$$

La sensibilità e la specificità del test diagnostico sono, rispettivamente:

$$\gamma = \Pr(Y = 1 \mid Z = 1) \quad \text{sensibilità;}$$

$$\delta = \Pr(Y = 0 \mid Z = 0) \quad \text{specificità.}$$

Le probabilità γ e δ sono entrambe indipendenti dalle variabili esplicative x nel modello e quindi l'errata classificazione della risposta non dipende da x.

Per un individuo per cui x è il vettore delle variabili esplicative, la regressione binomiale viene utilizzata per modellare la probabilità che tale individuo sperimenti lo stato di malattia D, ossia $\pi_x = \Pr(Z = 1 \mid x)$.

Nel caso della regressione logistica, $\pi_x = \exp(\beta^\top x)/(1 + \exp(\beta^\top x))$, dove β è un vettore di coefficienti di regressione. Alternativamente, $\pi_x = \Phi(\beta^\top x)$ per la regressione probit e $\pi_x = 1 - \exp(-\exp(\beta^\top x))$ per la regressione loglog complementare. Più in generale, $g(\pi_x) = \beta^\top x$, oppure $\pi_x = g^{-1}(\beta^\top x)$, per $g(\cdot)$ monotona, utilizzando la notazione dei modelli lineari generalizzati GLM, dove $g(\cdot)$ è la funzione legame. Dalla legge della probabilità totale, un risultato positivo del test per un individuo con variabile esplicativa x si presenta con probabilità

$$\begin{aligned} q_x &= \Pr(Y = 1 \mid x) \\ &= \Pr(Y = 1 \mid Z = 1) \Pr(Z = 1 \mid x) + \Pr(Y = 1 \mid Z = 0) \Pr(Z = 0 \mid x) \\ &= \gamma\pi_x + (1 - \delta)(1 - \pi_x) \\ &= \gamma\pi_x + 1 - \pi_x - \delta + \delta\pi_x \\ &= 1 - \delta - (1 - \gamma - \delta)\pi_x. \end{aligned}$$

Si osservano i dati (y_i, x_i), per $i = 1, ..., n$, dove y_i è il risultato del test diagnostico, mentre x_i è il vettore delle variabili esplicative per l'individuo i-esimo in un campione di numerosità n. Si assume $(Y_i \mid x_i) \sim$ Bernoulli(q_i), con probabilità di successo $q_i = q_{x_i}$ e $\pi_i = \pi_{x_i} = \Pr(Z_i = 1 \mid x_i)$. Si indica con X la matrice $(n \times k)$ delle variabili esplicative, e $y = (y_1, \ldots, y_n)$.
La funzione di verosimiglianza per β è

$$L(\beta) = \prod_{i=1}^{n} q_{x_i}^{y_i}(1 - q_{x_i})^{1-y_i},$$

dove

$$q_{x_i} = 1 - \delta - (1 - \gamma - \delta)\pi_{x_i}$$

e

$$1 - q_{x_i} = \delta + (1 - \gamma - \delta)\pi_{x_i}.$$

Quindi, la funzione di log-verosimiglianza per β è

$$l(\beta) = \sum_{i=1}^{n} y_i \log q_{x_i} + \sum_{i=1}^{n}(1 - y_i)\log(1 - q_{x_i}),$$

dove q_{x_i} è funzione di β, γ e δ. Nel seguito della trattazione γ e δ sono considerati come parametri fissati. Nel caso in cui γ e δ fossero invece ignoti, una possibile soluzione sarebbe quella di applicare i metodi bayesiani (cf. McInturff et al., 2004).

- Sia β un parametro scalare

La derivata della funzione di log-verosimiglianza rispetto a β è:

$$U_\beta = l_\beta = \frac{\partial}{\partial\beta} l(\beta).$$

Si ottiene la stima di massima verosimiglianza di β partendo dalla funzione di punteggio.
La funzione di punteggio per β è

$$U_\beta = \sum_{i=1}^{n} y_i \frac{\frac{\partial}{\partial\beta} q_{x_i}}{q_{x_i}} - \sum_{i=1}^{n}(1 - y_i)\frac{\frac{\partial}{\partial\beta} q_{x_i}}{1 - q_{x_i}},$$

che può essere scritta come

$$U_\beta = \sum_{i=1}^{n} y_i \frac{(\delta+\gamma-1)\frac{\partial \pi_{x_i}}{\partial \beta}}{1-\delta-(1-\gamma-\delta)\pi_{x_i}} - \sum_{i=1}^{n}(1-y_i)\frac{(\delta+\gamma-1)\frac{\partial \pi_{x_i}}{\partial \beta}}{\delta+(1-\gamma-\delta)\pi_{x_i}}.$$

Ponendo $U_\beta = 0$ si ottiene

$$\sum_{i=1}^{n} y_i \frac{(\delta+\gamma-1)\frac{\partial \pi_{x_i}}{\partial \beta}}{1-\delta-(1-\gamma-\delta)\pi_{x_i}} - \sum_{i=1}^{n}(1-y_i)\frac{(\delta+\gamma-1)\frac{\partial \pi_{x_i}}{\partial \beta}}{\delta+(1-\gamma-\delta)\pi_{x_i}} = 0,$$

equivalente a

$$\sum_{i=1}^{n} y_i \frac{(\delta+\gamma-1)\frac{\partial \pi_{x_i}}{\partial \beta}}{1-\delta-(1-\gamma-\delta)\pi_{x_i}} = \sum_{i=1}^{n}(1-y_i)\frac{(\delta+\gamma-1)\frac{\partial \pi_{x_i}}{\partial \beta}}{\delta+(1-\gamma-\delta)\pi_{x_i}},$$

da cui si ricava la stima di massima verosimiglianza di β, che è distorta per campioni finiti, ma asintoticamente non distorta.

Per ottenere l'informazione osservata $I(\beta) = -U_{\beta\beta}$ è necessario calcolare la derivata seconda della funzione di log-verosimiglianza $U_{\beta\beta} = \frac{\partial}{\partial \beta} U_\beta$:

$$\begin{aligned} U_{\beta\beta} &= \sum_{i=1}^{n} y_i \frac{\frac{\partial^2 q_{x_i}}{\partial\beta\partial\beta} q_{x_i} - \frac{\partial}{\partial\beta} q_{x_i} \frac{\partial}{\partial\beta} q_{x_i}}{q_{x_i}^2} - \sum_{i=1}^{n}(1-y_i)\frac{\frac{\partial^2 q_{x_i}}{\partial\beta\partial\beta}(1-q_{x_i)} + \frac{\partial}{\partial\beta} q_{x_i} \frac{\partial}{\partial\beta} q_{x_i}}{(1-q_{x_i})^2} \\ &= \sum_{i=1}^{n} y_i \frac{\frac{\partial^2 q_{x_i}}{\partial\beta\partial\beta} q_{x_i} - (\frac{\partial}{\partial\beta} q_{x_i})^2}{q_{x_i}^2} - \sum_{i=1}^{n}(1-y_i)\frac{\frac{\partial^2 q_{x_i}}{\partial\beta\partial\beta}(1-q_{x_i)} + (\frac{\partial}{\partial\beta} q_{x_i})^2}{(1-q_{x_i})^2}. \end{aligned}$$

- Sia β un parametro multidimensionale, in particolare $\beta = (\beta_1, \beta_2)$

Le derivate della funzione di log-verosimiglianza rispetto a β_1 e rispetto a β_2 sono:

$$U_{\beta_1} = l_{\beta_1} = \frac{\partial}{\partial \beta_1} l(\beta).$$

$$U_{\beta_2} = l_{\beta_2} = \frac{\partial}{\partial \beta_2} l(\beta).$$

Si ottengono le stime di massima verosimiglianza di β_1 e di β_2 partendo dalla funzione di punteggio.

La funzione di punteggio per β_1 è

$$U_{\beta_1} = \sum_{i=1}^{n} y_i \frac{\frac{\partial}{\partial \beta_1} q_{x_i}}{q_{x_i}} - \sum_{i=1}^{n} (1 - y_i) \frac{\frac{\partial}{\partial \beta_1} q_{x_i}}{1 - q_{x_i}},$$

mentre la funzione di punteggio per β_2 è

$$U_{\beta_2} = \sum_{i=1}^{n} y_i \frac{\frac{\partial}{\partial \beta_2} q_{x_i}}{q_{x_i}} - \sum_{i=1}^{n} (1 - y_i) \frac{\frac{\partial}{\partial \beta_2} q_{x_i}}{1 - q_{x_i}}.$$

Ponendo $U_{\beta_1} = 0$ e $U_{\beta_2} = 0$ si ottengono, rispettivamente,

$$\sum_{i=1}^{n} y_i \frac{\frac{\partial}{\partial \beta_1} q_{x_i}}{q_{x_i}} = \sum_{i=1}^{n} (1 - y_i) \frac{\frac{\partial}{\partial \beta_1} q_{x_i}}{1 - q_{x_i}}$$

e

$$\sum_{i=1}^{n} y_i \frac{\frac{\partial}{\partial \beta_2} q_{x_i}}{q_{x_i}} = \sum_{i=1}^{n} (1 - y_i) \frac{\frac{\partial}{\partial \beta_2} q_{x_i}}{1 - q_{x_i}},$$

da cui si ricavano le stime di massima verosimiglianza di β_1 e di β_2, che sono distorte per campioni finiti, ma asintoticamente non distorte.

Per ottenere la matrice di informazione osservata

$$I(\beta) = I(\beta_1, \beta_2) = - \begin{bmatrix} U_{\beta_1\beta_1} & U_{\beta_1\beta_2} \\ U_{\beta_2\beta_1} & U_{\beta_2\beta_2} \end{bmatrix}$$

è necessario calcolare le derivate seconde della log-verosimiglianza:

$$\begin{aligned} U_{\beta_1\beta_1} &= \frac{\partial}{\partial \beta_1} U_{\beta_1} \\ &= \sum_{i=1}^{n} y_i \frac{\frac{\partial^2 q_{x_i}}{\partial \beta_1^2} q_{x_i} - \left(\frac{\partial}{\partial \beta_1} q_{x_i} \right)^2}{q_{x_i}^2} - \sum_{i=1}^{n} (1 - y_i) \frac{\frac{\partial^2 q_{x_i}}{\partial \beta_1^2} (1 - q_{x_i}) + \left(\frac{\partial}{\partial \beta_1} q_{x_i} \right)^2}{(1 - q_{x_i})^2}; \end{aligned}$$

$$\begin{aligned} U_{\beta_2\beta_2} &= \frac{\partial}{\partial \beta_2} U_{\beta_2} \\ &= \sum_{i=1}^{n} y_i \frac{\frac{\partial^2 q_{x_i}}{\partial \beta_2^2} q_{x_i} - \left(\frac{\partial}{\partial \beta_2} q_{x_i} \right)^2}{q_{x_i}^2} - \sum_{i=1}^{n} (1 - y_i) \frac{\frac{\partial^2 q_{x_i}}{\partial \beta_2^2} (1 - q_{x_i}) + \left(\frac{\partial}{\partial \beta_2} q_{x_i} \right)^2}{(1 - q_{x_i})^2}; \end{aligned}$$

$$\begin{aligned} U_{\beta_1\beta_2} &= \frac{\partial}{\partial\beta_2}U_{\beta_1} = \frac{\partial}{\partial\beta_1}U_{\beta_2} = U_{\beta_2\beta_1} \\ &= \sum_{i=1}^{n} y_i \frac{\frac{\partial^2 q_{x_i}}{\partial\beta_1\beta_2}q_{x_i} - \left(\frac{\partial}{\partial\beta_2}q_{x_i}\right)\left(\frac{\partial}{\partial\beta_1}q_{x_i}\right)}{q_{x_i}^2} \\ &\quad - \sum_{i=1}^{n}(1-y_i)\frac{\frac{\partial^2 q_{x_i}}{\partial\beta_1\beta_2}(1-q_{x_i)} + \left(\frac{\partial}{\partial\beta_2}q_{x_i}\right)\left(\frac{\partial}{\partial\beta_1}q_{x_i}\right)}{(1-q_{x_i})^2}. \end{aligned}$$

2.5 Correzione della distorsione con il metodo di Firth

Nel caso in cui β sia un parametro scalare, la funzione di punteggio modificata secondo il metodo di Firth (1993), cfr. paragrafi 1.7 e 1.9, può essere scritta come

$$U_\beta^* = U_\beta + A_\beta.$$

A_β è basato sull'informazione attesa, nel qual caso si scrive $A_\beta = A_\beta^{(E)}$, oppure sull'informazione osservata, nel qual caso si scrive invece $A_\beta = A_\beta^{(O)}$. Le stime derivate utilizzando $A_\beta^{(O)}$ risultano preferibili in termini di efficienza, come richiamato nel paragrafo 1.9.
Sia quindi A_β basato sull'informazione osservata, per cui

$$A_\beta = A_\beta^{(O)} = -U_{\beta s}\kappa^{s,t}\kappa^{u,v}\frac{(\kappa_{t,u,v} + \kappa_{t,uv})}{2n}.$$

Dal momento che γ e δ sono considerati come parametri fissi,

$$A_\beta = -U_{\beta\beta}\kappa^{\beta,\beta}\kappa^{\beta,\beta}\frac{(\kappa_{\beta,\beta,\beta} + \kappa_{\beta,\beta\beta})}{2n} = -\frac{U_{\beta\beta}}{2n}(\kappa^{\beta,\beta})^2(\kappa_{\beta,\beta,\beta} + \kappa_{\beta,\beta\beta}),$$

dove

$$\kappa_{\beta,\beta} = \frac{1}{n}E\{U_\beta U_\beta\} = \frac{1}{n}E\{U_\beta^2\},$$

$$\kappa_{\beta,\beta,\beta} = \frac{1}{n}E\{U_\beta U_\beta U_\beta\} = \frac{1}{n}E\{U_\beta^3\},$$

$$\kappa_{\beta,\beta\beta} = \frac{1}{n}E\{U_\beta U_{\beta\beta}\}.$$

Si scrivano le derivate prima e seconda della log-verosimiglianza rispettivamente come

$$\begin{aligned} U_\beta &= \sum_{i=1}^{n} y_i a_i - \sum_{i=1}^{n} (1-y_i) b_i \\ &= \sum_{i=1}^{n} y_i a_i - \sum_{i=1}^{n} b_i + \sum_{i=1}^{n} y_i b_i \\ &= \sum_{i=1}^{n} y_i (a_i + b_i) - \sum_{i=1}^{n} b_i \end{aligned}$$

e

$$\begin{aligned} U_{\beta\beta} &= \sum_{i=1}^{n} y_i c_i - \sum_{i=1}^{n} (1-y_i) d_i \\ &= \sum_{i=1}^{n} y_i c_i - \sum_{i=1}^{n} d_i + \sum_{i=1}^{n} y_i d_i \\ &= \sum_{i=1}^{n} y_i (c_i + d_i) - \sum_{i=1}^{n} d_i, \end{aligned}$$

dove

$$a_i = \frac{\frac{\partial}{\partial\beta} q_{x_i}}{q_{x_i}},$$

$$b_i = \frac{\frac{\partial}{\partial\beta} q_{x_i}}{1 - q_{x_i}},$$

$$c_i = \frac{\frac{\partial^2 q_{x_i}}{\partial\beta\partial\beta} q_{x_i} - (\frac{\partial}{\partial\beta} q_{x_i})^2}{q_{x_i}^2},$$

$$d_i = \frac{\frac{\partial^2 q_{x_i}}{\partial\beta\partial\beta} (1 - q_{x_i)} + (\frac{\partial}{\partial\beta} q_{x_i})^2}{(1 - q_{x_i})^2}.$$

A questo punto, si possono calcolare tutte le quantità necessarie per determinare A_β.

- $\kappa_{\beta,\beta} = \frac{1}{n} E\{U_\beta U_\beta\} = \frac{1}{n} E\{U_\beta^2\}$

Poiché $(Y_i \mid x_i) \sim$ Bernoulli(q_i), allora $E(Y_i) = q_i$ e $Var(Y_i) = q_i(1-q_i)$. Si ha dunque:

$$\begin{aligned} Var(U_\beta) &= \sum_{i=1}^{n}(a_i+b_i)^2 Var(y_i) \\ &= \sum_{i=1}^{n}(a_i+b_i)^2 q_i(1-q_i) \end{aligned}$$

e quindi, ricordando che la media della *score* è zero, risulta

$$E(U_\beta^2) = Var(U_\beta) + [E(U_\beta)]^2 = \sum_{i=1}^{n}(a_i+b_i)^2 q_i(1-q_i).$$

- $\kappa_{\beta,\beta,\beta} = \frac{1}{n}E\{U_\beta U_\beta U_\beta\} = \frac{1}{n}E\{U_\beta^3\}$

$$E(U_\beta^3) = \kappa_3(U_\beta) - 3E(U_\beta)E(U_\beta^2) + 2[E(U_\beta)]^3 = \kappa_3(U_\beta),$$

dove κ_3 indica il cumulante terzo di U_β. In particolare, sfruttando la proprietà di invarianza rispetto a traslazione del cumulante terzo e poiché $(a_i + b_i) > 0$, si ha che

$$\begin{aligned} \kappa_3(U_\beta) &= \kappa_3\Big(\sum_{i=1}^{n} y_i(a_i+b_i) - \sum_{i=1}^{n} b_i\Big) \\ &= \sum_{i=1}^{n}\kappa_3(y_i(a_i+b_i)) \\ &= \sum_{i=1}^{n}\kappa_3(y_i)(a_i+b_i)^3. \end{aligned}$$

In generale, per le famiglie esponenziali

$$p(y;\theta) = \exp\{\theta y - k(\theta)\}c(y),$$

$$\kappa_r = K^{(r)}(\theta),$$

ossia il cumulante di ordine r è uguale alla derivata r-sima di K in θ.
Se, ad esempio, $Y \sim$ Bernoulli(π), allora

$$\begin{aligned} p(y;\pi) &= \pi^y(1-\pi)^{1-y} \\ &= \left(\frac{\pi}{1-\pi}\right)^y (1-\pi) \\ &= \exp\left\{y\log\frac{\pi}{1-\pi} + \log(1-\pi)\right\}. \end{aligned}$$

Ponendo

$$\theta = \log \frac{\pi}{1-\pi},$$

si ottiene

$$e^{\theta} = e^{\log \frac{\pi}{1-pi}} = \frac{\pi}{1-\pi},$$

da cui

$$e^{\theta} - \pi e^{\theta} = \pi$$

e quindi

$$\pi = \frac{e^{\theta}}{1+e^{\theta}} \qquad \text{e} \qquad 1-\pi = \frac{1}{1+e^{\theta}}.$$

Si può dunque scrivere

$$\begin{aligned} p(y;\pi) &= \exp\left\{\theta y + \log\frac{1}{1+e^{\theta}}\right\} \\ &= \exp\left\{\theta y + \log 1 - \log(1+e^{\theta})\right\} \\ &= \exp\{\theta y - \underbrace{\log(1+e^{\theta})}_{K(\theta)}\}. \end{aligned}$$

$$K(\theta) = \log(1+e^{\theta});$$

$$K'(\theta) = E(Y) = \frac{e^{\theta}}{1+e^{\theta}} = \pi;$$

$$\begin{aligned} K''(\theta) &= Var(Y) = \frac{e^{\theta}(1+e^{\theta}) - e^{\theta}e^{\theta}}{(1+e^{\theta})^2} \\ &= \frac{e^{\theta}}{(1+e^{\theta})^2} = \frac{e^{\theta}}{(1+e^{\theta})}\frac{1}{(1+e^{\theta})} = \pi(1-\pi); \end{aligned}$$

$$
\begin{aligned}
K'''(\theta) &= \frac{e^\theta(1+e^\theta)^2 - 2(1+e^\theta)e^\theta e^\theta}{(1+e^\theta)^4} \\
&= \frac{e^\theta(1+e^{2\theta}+2e^\theta) - 2e^{2\theta} - 2e^{3\theta}}{(1+e^\theta)^4} \\
&= \frac{e^\theta + e^{3\theta} + 2e^{2\theta} - 2e^{2\theta} - 2e^{3\theta}}{(1+e^\theta)^4} \\
&= \frac{e^\theta - e^{3\theta}}{(1+e^\theta)^4} \\
&= \frac{e^\theta}{(1+e^\theta)^4} - \frac{e^{3\theta}}{(1+e^\theta)^4} \\
&= \frac{e^\theta}{(1+e^\theta)}\frac{1}{(1+e^\theta)^3} - \frac{e^{3\theta}}{(1+e^\theta)^3}\frac{1}{(1+e^\theta)} \\
&= \pi\left(\frac{1}{1+e^\theta}\right)^3 - \left(\frac{e^\theta}{1+e^\theta}\right)^3(1-\pi) \\
&= \pi(1-\pi)^3 - \pi^3(1-\pi) \\
&= \pi(1-\pi)[(1-\pi)^2 - \pi^2] = \pi(1-\pi)(1-2\pi) = \kappa_3(Y).
\end{aligned}
$$

Nel caso in esame, $(Y_i \mid x_i) \sim$ Bernoulli(q_i), quindi $\kappa_3(y_i) = q_i(1-q_i)(1-2q_i)$ e $\kappa_3(U_\beta) = \sum_{i=1}^n q_i(1-q_i)(1-2q_i)(a_i+b_i)^3$.

- $\kappa_{\beta,\beta\beta} = \frac{1}{n}E\{U_\beta U_{\beta\beta}\}$

$$
\begin{aligned}
U_\beta U_{\beta\beta} &= \left(\sum_{i=1}^n y_i(a_i+b_i) - \sum_{i=1}^n b_i\right)\left(\sum_{j=1}^n y_j(c_j+d_j) - \sum_{j=1}^n d_j\right) \\
&= \underbrace{\sum_{i=1}^n\sum_{j=1}^n y_i(a_i+b_i)y_j(c_j+d_j)}_{A} - \underbrace{\sum_{i=1}^n\sum_{j=1}^n y_i(a_i+b_i)d_j}_{B} \\
&\quad - \underbrace{\sum_{i=1}^n\sum_{j=1}^n y_j(c_j+d_j)b_i}_{C} + \underbrace{\sum_{i=1}^n\sum_{j=1}^n b_i d_j}_{D}
\end{aligned}
$$

In particolare,

$$\begin{aligned} A &= \sum_{i=1}^{n}\sum_{j=1}^{n} y_i(a_i+b_i)y_j(c_j+d_j) \\ &= \sum_{i=j} y_i^2(a_i+b_i)(c_i+d_i) + \sum_{i\neq j} y_i(a_i+b_i)y_j(c_j+d_j), \end{aligned}$$

$$\begin{aligned} B &= \sum_{i=1}^{n}\sum_{j=1}^{n} y_i(a_i+b_i)d_j \\ &= \sum_{i=j} y_i(a_i+b_i)d_i + \sum_{i\neq j} y_i(a_i+b_i)d_j, \end{aligned}$$

$$\begin{aligned} C &= \sum_{i=1}^{n}\sum_{j=1}^{n} y_j(c_j+d_j)b_i \\ &= \sum_{i=j} y_i(c_i+d_i)b_i + \sum_{i\neq j} y_j(c_j+d_j)b_i, \end{aligned}$$

$$\begin{aligned} D &= \sum_{i=1}^{n}\sum_{j=1}^{n} b_i d_j \\ &= \sum_{i=j} b_i d_i + \sum_{i\neq j} b_i d_j. \end{aligned}$$

Ricordando che $E(Y_i) = q_i$ e $E(Y_i^2) = q_i$, si può scrivere

$$E(A) = \sum_{i=j} q_i(a_i+b_i)(c_i+d_i) + \sum_{i\neq j} q_i(a_i+b_i)q_j(c_j+d_j),$$

$$E(B) = \sum_{i=j} q_i(a_i+b_i)d_i + \sum_{i\neq j} q_i(a_i+b_i)d_j,$$

$$E(C) = \sum_{i=j} q_i(c_i+d_i)b_i + \sum_{i\neq j} q_j(c_j+d_j)b_i,$$

$$E(D) = \sum_{i=j} b_i d_i + \sum_{i\neq j} b_i d_j.$$

Di conseguenza,

$$\kappa_{\beta,\beta\beta} = \frac{1}{n}E\{U_\beta U_{\beta\beta}\} = \frac{1}{n}\left\{E(A) - E(B) - E(C) + E(D)\right\}.$$

Nel caso in cui β sia un parametro multidimensionale, $\beta = (\beta_1, \beta_2)$, la funzione di punteggio modificata può essere scritta come

$$U^*_\beta = U_\beta + A_\beta = \begin{pmatrix} U_{\beta_1} + A_{\beta_1} \\ U_{\beta_2} + A_{\beta_2} \end{pmatrix}.$$

Siano γ e δ considerati come parametri fissi e sia $A_\beta = (A_{\beta_1}, A_{\beta_2})$ basato sull'informazione osservata. Risulta quindi

$$\begin{aligned} A_{\beta_1} &= -U_{\beta_1 s}\kappa^{s,t}\kappa^{u,v}\frac{(\kappa_{t,u,v} + \kappa_{t,uv})}{2n} \\ &= -\frac{1}{2n}\sum_{s=1}^{2}\sum_{t=1}^{2}\sum_{u=1}^{2}\sum_{v=1}^{2} U_{\beta_1 s}\kappa^{s,t}\kappa^{u,v}(\kappa_{t,u,v} + \kappa_{t,uv}) \end{aligned}$$

e

$$\begin{aligned} A_{\beta_2} &= -U_{\beta_2 s}\kappa^{s,t}\kappa^{u,v}\frac{(\kappa_{t,u,v} + \kappa_{t,uv})}{2n} \\ &= -\frac{1}{2n}\sum_{s=1}^{2}\sum_{t=1}^{2}\sum_{u=1}^{2}\sum_{v=1}^{2} U_{\beta_2 s}\kappa^{s,t}\kappa^{u,v}(\kappa_{t,u,v} + \kappa_{t,uv}), \end{aligned}$$

dove $s, t, u, v = 1$ indica $s, t, u, v = \beta_1$ e $s, t, u, v = 2$ indica $s, t, u, v = \beta_2$. Si scrivano le derivate prime e seconde della log-verosimiglianza come

$$U_{\beta_1} = \sum_{i=1}^{n} y_i(a_{1i} + b_{1i}) - \sum_{i=1}^{n} b_{1i},$$

$$U_{\beta_2} = \sum_{i=1}^{n} y_i(a_{2i} + b_{2i}) - \sum_{i=1}^{n} b_{2i},$$

$$U_{\beta_1\beta_1} = \sum_{i=1}^{n} y_i(c_{1i} + d_{1i}) - \sum_{i=1}^{n} d_{1i},$$

$$U_{\beta_2\beta_2} = \sum_{i=1}^{n} y_i(c_{2i} + d_{2i}) - \sum_{i=1}^{n} d_{2i},$$

$$U_{\beta_1\beta_2} = U_{\beta_2\beta_1} = \sum_{i=1}^{n} y_i(c_{12i} + d_{12i}) - \sum_{i=1}^{n} d_{12i},$$

dove

$$a_{1i} = \frac{\frac{\partial}{\partial\beta_1}q_{x_i}}{q_{x_i}}, \quad b_{1i} = \frac{\frac{\partial}{\partial\beta_1}q_{x_i}}{1-q_{x_i}}, \quad a_{2i} = \frac{\frac{\partial}{\partial\beta_2}q_{x_i}}{q_{x_i}}, \quad b_{2i} = \frac{\frac{\partial}{\partial\beta_2}q_{x_i}}{1-q_{x_i}},$$

$$c_{1i} = \frac{\frac{\partial^2 q_{x_i}}{\partial\beta_1^2}q_{x_i} - \left(\frac{\partial}{\partial\beta_1}q_{x_i}\right)^2}{q_{x_i}^2}, \quad d_{1i} = \frac{\frac{\partial^2 q_{x_i}}{\partial\beta_1^2}(1-q_{x_i}) + \left(\frac{\partial}{\partial\beta_1}q_{x_i}\right)^2}{(1-q_{x_i})^2},$$

$$c_{2i} = \frac{\frac{\partial^2 q_{x_i}}{\partial\beta_2^2}q_{x_i} - \left(\frac{\partial}{\partial\beta_2}q_{x_i}\right)^2}{q_{x_i}^2}, \quad d_{2i} = \frac{\frac{\partial^2 q_{x_i}}{\partial\beta_2^2}(1-q_{x_i}) + \left(\frac{\partial}{\partial\beta_2}q_{x_i}\right)^2}{(1-q_{x_i})^2},$$

$$c_{12i} = \frac{\frac{\partial^2 q_{x_i}}{\partial\beta_1\beta_2}q_{x_i} - \frac{\partial}{\partial\beta_1}q_{x_i}\frac{\partial}{\partial\beta_2}q_{x_i}}{q_{x_i}^2}, \quad d_{12i} = \frac{\frac{\partial^2 q_{x_i}}{\partial\beta_1\beta_2}(1-q_{x_i}) + \frac{\partial}{\partial\beta_1}q_{x_i}\frac{\partial}{\partial\beta_2}q_{x_i}}{(1-q_{x_i})^2}.$$

A questo punto si possono calcolare tutte le quantità necessarie per determinare A_{β_1} e A_{β_2}.

- $\kappa_{\beta_1,\beta_1} = \frac{1}{n}E\{U_{\beta_1}^2\} = \frac{1}{n}\sum_{i=1}^n E[(U_{\beta_1}^i)^2]$

$$U_{\beta_1}^i = (a_{1i}+b_{1i})y_i - b_{1i},$$

$$(U_{\beta_1}^i)^2 = (a_{1i}+b_{1i})^2 y_i^2 + b_{1i}^2 - 2(a_{1i}+b_{1i})b_{1i}y_i,$$

quindi

$$E[(U_{\beta_1}^i)^2] = (a_{1i}+b_{1i})^2 q_{x_i} + b_{1i}^2 - 2(a_{1i}+b_{1i})b_{1i}q_{x_i}.$$

- $\kappa_{\beta_2,\beta_2} = \frac{1}{n}E\{U_{\beta_2}^2\} = \frac{1}{n}\sum_{i=1}^n E[(U_{\beta_1}^i)^2]$

$$U_{\beta_2}^i = (a_{2i}+b_{2i})y_i - b_{2i},$$

$$(U_{\beta_2}^i)^2 = (a_{2i}+b_{2i})^2 y_i^2 + b_{2i}^2 - 2(a_{2i}+b_{2i})b_{2i}y_i,$$

quindi

$$E[(U_{\beta_2}^i)^2] = (a_{2i}+b_{2i})^2 q_{x_i} + b_{2i}^2 - 2(a_{2i}+b_{2i})b_{2i}q_{x_i}.$$

- $\kappa_{\beta_1,\beta_2} = \frac{1}{n}E\{U_{\beta_1}U_{\beta_2}\} = \frac{1}{n}\sum_{i=1}^n E[U_{\beta_1}^i U_{\beta_2}^i] = \kappa_{\beta_2,\beta_1}$

$$U_{\beta_1}^i = (a_{1i}+b_{1i})y_i - b_{1i},$$

$$U^i_{\beta_2} = (a_{2i} + b_{2i})y_i - b_{2i},$$

$$U^i_{\beta_1} U^i_{\beta_2} = (a_{1i} + b_{1i})(a_{2i} + b_{2i})y_i^2 - (a_{1i} + b_{1i})b_{2i}y_i - (a_{2i} + b_{2i})b_{1i}y_i + b_{1i}b_{2i},$$

quindi

$$E[U^i_{\beta_1} U^i_{\beta_2}] = (a_{1i}+b_{1i})(a_{2i}+b_{2i})q_{x_i} - (a_{1i}+b_{1i})b_{2i}q_{x_i} - (a_{2i}+b_{2i})b_{1i}q_{x_i} + b_{1i}b_{2i}.$$

- $\kappa_{\beta_1,\beta_1,\beta_1} = \frac{1}{n} E\{U^3_{\beta_1}\} = \frac{1}{n}\sum_{i=1}^n E[(U^i_{\beta_1})^3]$

$$U^i_{\beta_1} = (a_{1i} + b_{1i})y_i - b_{1i},$$

$$(U^i_{\beta_1})^3 = (a_{1i} + b_{1i})^3 y_i^3 - b_{1i}^3 - 3(a_{1i} + b_{1i})^2 b_{1i} y_i^2 + 3(a_{1i} + b_{1i})b_{1i}^2 y_i,$$

quindi

$$E[(U^i_{\beta_1})^3] = (a_{1i} + b_{1i})^3 q_{x_i} - b_{1i}^3 - 3(a_{1i} + b_{1i})^2 b_{1i} q_{x_i} + 3(a_{1i} + b_{1i})b_{1i}^2 q_{x_i}.$$

- $\kappa_{\beta_2,\beta_2,\beta_2} = \frac{1}{n} E\{U^3_{\beta_2}\} = \frac{1}{n}\sum_{i=1}^n E[(U^i_{\beta_2})^3]$

$$U^i_{\beta_2} = (a_{2i} + b_{2i})y_i - b_{2i},$$

$$(U^i_{\beta_2})^3 = (a_{2i} + b_{2i})^3 y_i^3 - b_{2i}^3 - 3(a_{2i} + b_{2i})^2 b_{2i} y_i^2 + 3(a_{2i} + b_{2i})b_{2i}^2 y_i,$$

quindi

$$E[(U^i_{\beta_2})^3] = (a_{2i} + b_{2i})^3 q_{x_i} - b_{2i}^3 - 3(a_{2i} + b_{2i})^2 b_{2i} q_{x_i} + 3(a_{2i} + b_{2i})b_{2i}^2 q_{x_i}.$$

- $\kappa_{\beta_1,\beta_1\beta_1} = \frac{1}{n} E\{U_{\beta_1} U_{\beta_1\beta_1}\} = \frac{1}{n}\sum_{i=1}^n E[U^i_{\beta_1} U^i_{\beta_1\beta_1}]$

$$U^i_{\beta_1} = (a_{1i} + b_{1i})y_i - b_{1i},$$

$$U^i_{\beta_1\beta_1} = (c_{1i} + d_{1i})y_i - d_{1i},$$

$$U^i_{\beta_1} U^i_{\beta_1\beta_1} = (a_{1i} + b_{1i})(c_{1i} + d_{1i})y_i^2 - (a_{1i} + b_{1i})d_{1i}y_i - (c_{1i} + d_{1i})b_{1i}y_i + b_{1i}d_{1i},$$

quindi

$$E[U^i_{\beta_1} U^i_{\beta_1\beta_1}] = (a_{1i}+b_{1i})(c_{1i}+d_{1i})q_{x_i} - (a_{1i}+b_{1i})d_{1i}q_{x_i} - (c_{1i}+d_{1i})b_{1i}q_{x_i} + b_{1i}d_{1i}.$$

- $\kappa_{\beta_2,\beta_2\beta_2} = \frac{1}{n}E\{U_{\beta_2}U_{\beta_2\beta_2}\} = \frac{1}{n}\sum_{i=1}^{n} E[U^i_{\beta_2}U^i_{\beta_2\beta_2}]$

$$U^i_{\beta_2} = (a_{2i} + b_{2i})y_i - b_{2i},$$

$$U^i_{\beta_2\beta_2} = (c_{2i} + d_{2i})y_i - d_{2i},$$

$U^i_{\beta_2}U^i_{\beta_2\beta_2} = (a_{2i} + b_{2i})(c_{2i} + d_{2i})y_i^2 - (a_{2i} + b_{2i})d_{2i}y_i - (c_{2i} + d_{2i})b_{2i}y_i + b_{2i}d_{2i}$,

quindi

$E[U^i_{\beta_2}U^i_{\beta_2\beta_2}] = (a_{2i}+b_{2i})(c_{2i}+d_{2i})q_{x_i} - (a_{2i}+b_{2i})d_{2i}q_{x_i} - (c_{2i}+d_{2i})b_{2i}q_{x_i} + b_{2i}d_{2i}$.

- $\kappa_{\beta_1,\beta_1\beta_2} = \frac{1}{n}E\{U_{\beta_1}U_{\beta_1\beta_2}\} = \frac{1}{n}\sum_{i=1}^{n} E[U^i_{\beta_1}U^i_{\beta_1\beta_2}] = \kappa_{\beta_1,\beta_2\beta_1}$

$$U^i_{\beta_1} = (a_{1i} + b_{1i})y_i - b_{1i},$$

$$U^i_{\beta_1\beta_2} = (c_{12i} + d_{12i})y_i - d_{12i},$$

$U^i_{\beta_1}U^i_{\beta_1\beta_2} = (a_{1i}+b_{1i})(c_{12i}+d_{12i})y_i^2 - (a_{1i}+b_{1i})d_{12i}y_i - (c_{12i}+d_{12i})b_{1i}y_i + b_{1i}d_{12i}$,

quindi

$E[U^i_{\beta_1}U^i_{\beta_1\beta_2}] = (a_{1i}+b_{1i})(c_{12i}+d_{12i})q_{x_i} - (a_{1i}+b_{1i})d_{12i}q_{x_i} - (c_{12i}+d_{12i})b_{1i}q_{x_i} + b_{1i}d_{12i}$.

- $\kappa_{\beta_1,\beta_2\beta_2} = \frac{1}{n}E\{U_{\beta_1}U_{\beta_2\beta_2}\} = \frac{1}{n}\sum_{i=1}^{n} E[U^i_{\beta_1}U^i_{\beta_2\beta_2}]$

$$U^i_{\beta_1} = (a_{1i} + b_{1i})y_i - b_{1i},$$

$$U^i_{\beta_2\beta_2} = (c_{2i} + d_{2i})y_i - d_{2i},$$

$U^i_{\beta_1}U^i_{\beta_2\beta_2} = (a_{1i} + b_{1i})(c_{2i} + d_{2i})y_i^2 - (a_{1i} + b_{1i})d_{2i}y_i - (c_{2i} + d_{2i})b_{1i}y_i + b_{1i}d_{2i}$,

quindi

$E[U^i_{\beta_1}U^i_{\beta_2\beta_2}] = (a_{1i}+b_{1i})(c_{2i}+d_{2i})q_{x_i} - (a_{1i}+b_{1i})d_{2i}q_{x_i} - (c_{2i}+d_{2i})b_{1i}q_{x_i} + b_{1i}d_{2i}$.

- $\kappa_{\beta_2,\beta_1\beta_1} = \frac{1}{n}E\{U_{\beta_2}U_{\beta_1\beta_1}\} = \frac{1}{n}\sum_{i=1}^{n} E[U^i_{\beta_2}U^i_{\beta_1\beta_1}]$

$$U^i_{\beta_2} = (a_{2i} + b_{2i})y_i - b_{2i},$$

$$U^i_{\beta_1\beta_1} = (c_{1i} + d_{1i})y_i - d_{1i},$$

$$U^i_{\beta_2}U^i_{\beta_1\beta_1} = (a_{2i} + b_{2i})(c_{1i} + d_{1i})y_i^2 - (a_{2i} + b_{2i})d_{1i}y_i - (c_{1i} + d_{1i})b_{2i}y_i + b_{2i}d_{1i},$$

quindi

$$E[U^i_{\beta_2}U^i_{\beta_1\beta_1}] = (a_{2i}+b_{2i})(c_{1i}+d_{1i})q_{x_i} - (a_{2i}+b_{2i})d_{1i}q_{x_i} - (c_{1i}+d_{1i})b_{2i}q_{x_i} + b_{2i}d_{1i}.$$

- $\kappa_{\beta_2,\beta_1\beta_2} = \frac{1}{n}E\{U_{\beta_2}U_{\beta_1\beta_2}\} = \frac{1}{n}\sum_{i=1}^n E[U^i_{\beta_2}U^i_{\beta_1\beta_2}] = \kappa_{\beta_2,\beta_2\beta_1}$

$$U^i_{\beta_2} = (a_{2i} + b_{2i})y_i - b_{2i},$$

$$U^i_{\beta_1\beta_2} = (c_{12i} + d_{12i})y_i - d_{12i},$$

$$U^i_{\beta_2}U^i_{\beta_1\beta_2} = (a_{2i}+b_{2i})(c_{12i}+d_{12i})y_i^2 - (a_{2i}+b_{2i})d_{12i}y_i - (c_{12i}+d_{12i})b_{2i}y_i + b_{2i}d_{12i},$$

quindi

$$E[U^i_{\beta_2}U^i_{\beta_1\beta_2}] = (a_{2i}+b_{2i})(c_{12i}+d_{12i})q_{x_i} - (a_{2i}+b_{2i})d_{12i}q_{x_i} - (c_{12i}+d_{12i})b_{2i}q_{x_i} + b_{2i}d_{12i}.$$

- $\kappa_{\beta_1,\beta_1,\beta_2} = \frac{1}{n}E\{U^2_{\beta_1}U_{\beta_2}\} = \frac{1}{n}\sum_{i=1}^n E[(U^i_{\beta_1})^2U^i_{\beta_2}] = \kappa_{\beta_1,\beta_2,\beta_1} = \kappa_{\beta_2,\beta_1,\beta_1}$

$$U^i_{\beta_1} = (a_{1i} + b_{1i})y_i - b_{1i},$$

$$U^i_{\beta_2} = (a_{2i} + b_{2i})y_i - b_{2i},$$

$$(U^i_{\beta_1})^2 = (a_{1i} + b_{1i})^2y_i^2 + b_{1i}^2 - 2(a_{1i} + b_{1i})b_{1i}y_i,$$

$$\begin{aligned}(U^i_{\beta_1})^2U^i_{\beta_2} &= (a_{1i} + b_{1i})^2(a_{2i} + b_{2i})y_i^3 - (a_{1i} + b_{1i})^2b_{2i}y_i^2 + (a_{2i} + b_{2i})b_{1i}^2y_i \\ &\quad -b_{1i}^2b_{2i} - 2(a_{1i} + b_{1i})b_{1i}(a_{2i} + b_{2i})y_i^2 + 2(a_{1i} + b_{1i})b_{1i}b_{2i}y_i,\end{aligned}$$

quindi

$$\begin{aligned}E[(U^i_{\beta_1})^2U^i_{\beta_2}] &= (a_{1i} + b_{1i})^2(a_{2i} + b_{2i})q_{x_i} - (a_{1i} + b_{1i})^2b_{2i}q_{x_i} \\ &\quad +(a_{2i} + b_{2i})b_{1i}^2q_{x_i} - b_{1i}^2b_{2i} - 2(a_{1i} + b_{1i})b_{1i}(a_{2i} + b_{2i})q_{x_i} \\ &\quad +2(a_{1i} + b_{1i})b_{1i}b_{2i}q_{x_i}.\end{aligned}$$

- $\kappa_{\beta_1,\beta_2,\beta_2} = \frac{1}{n}E\{U_{\beta_1}U^2_{\beta_2}\} = \frac{1}{n}\sum_{i=1}^{n} E[U^i_{\beta_1}(U^i_{\beta_2})^2] = \kappa_{\beta_2,\beta_1,\beta_2} = \kappa_{\beta_2,\beta_2,\beta_1}$

$$U^i_{\beta_1} = (a_{1i} + b_{1i})y_i - b_{1i},$$

$$U^i_{\beta_2} = (a_{2i} + b_{2i})y_i - b_{2i},$$

$$(U^i_{\beta_2})^2 = (a_{2i} + b_{2i})^2 y_i^2 + b_{2i}^2 - 2(a_{2i} + b_{2i})b_{2i}y_i,$$

$$\begin{aligned} U^i_{\beta_1}(U^i_{\beta_2})^2 &= (a_{1i} + b_{1i})(a_{2i} + b_{2i})^2 y_i^3 - (a_{2i} + b_{2i})^2 b_{1i} y_i^2 + (a_{1i} + b_{1i})b_{2i}^2 y_i \\ &\quad -b_{1i}b_{2i}^2 - 2(a_{2i} + b_{2i})b_{2i}(a_{1i} + b_{1i})y_i^2 + 2(a_{2i} + b_{2i})b_{1i}b_{2i}y_i, \end{aligned}$$

quindi

$$\begin{aligned} E[U^i_{\beta_1}(U^i_{\beta_2})^2] &= (a_{1i} + b_{1i})(a_{2i} + b_{2i})^2 q_{x_i} - (a_{2i} + b_{2i})^2 b_{1i} q_{x_i} \\ &\quad +(a_{1i} + b_{1i})b_{2i}^2 q_{x_i} - b_{1i}b_{2i}^2 - 2(a_{2i} + b_{2i})b_{2i}(a_{1i} + b_{1i})q_{x_i} \\ &\quad +2(a_{2i} + b_{2i})b_{1i}b_{2i}q_{x_i}. \end{aligned}$$

2.6 Casi particolari del modello

2.6.1 Il modello logit

Si consideri un modello logit con una variabile esplicativa scalare e intercetta zero. Si ha

$$\pi_{x_i} = \frac{e^{\beta x_i}}{1 + e^{\beta x_i}}.$$

Risulta quindi

$$\begin{aligned} \frac{\partial}{\partial\beta}q_{x_i} &= -(1 - \gamma - \delta)\frac{\partial}{\partial\beta}\pi_{x_i} \\ &= (\delta + \gamma - 1)\frac{e^{\beta x_i}x_i(1 + e^{\beta x_i}) - e^{\beta x_i}e^{\beta x_i}x_i}{(1 + e^{\beta x_i})^2} \\ &= (\delta + \gamma - 1)\frac{e^{\beta x_i}x_i(1 + e^{\beta x_i} - e^{\beta x_i})}{(1 + e^{\beta x_i})^2} \\ &= (\delta + \gamma - 1)\frac{x_i e^{\beta x_i}}{(1 + e^{\beta x_i})^2} \end{aligned}$$

e

$$\begin{aligned}
\frac{\partial^2 q_{x_i}}{\partial \beta^2} &= -(1-\gamma-\delta)\frac{x_i^2 e^{\beta x_i}(1+e^{\beta x_i})^2 - x_i e^{\beta x_i} 2(1+e^{\beta x_i}) x_i e^{\beta x_i}}{(1+e^{\beta x_i})^4} \\
&= -(1-\gamma-\delta)\frac{x_i^2 e^{\beta x_i}(1+e^{\beta x_i})^2 - 2x_i^2 e^{2\beta x_i}(1+e^{\beta x_i})}{(1+e^{\beta x_i})^4} \\
&= -(1-\gamma-\delta)\frac{x_i^2 e^{\beta x_i}(1+e^{\beta x_i} - 2e^{\beta x_i})}{(1+e^{\beta x_i})^3} \\
&= -(1-\gamma-\delta)\frac{x_i^2 e^{\beta x_i}(1-e^{\beta x_i})}{(1+e^{\beta x_i})^3}.
\end{aligned}$$

Introducendo un parametro di intercetta, si ha $\beta = (\beta_1, \beta_2)$ e

$$\pi_{x_i} = \frac{e^{\beta_1+\beta_2 x_i}}{1+e^{\beta_1+\beta_2 x_i}}.$$

Si ha quindi

$$\begin{aligned}
\frac{\partial}{\partial \beta_1} q_{x_i} &= -(1-\gamma-\delta)\frac{\partial}{\partial \beta_1}\pi_{x_i} \\
&= -(1-\gamma-\delta)\frac{e^{\beta_1+\beta_2 x_i}(1+e^{\beta_1+\beta_2 x_i}) - e^{\beta_1+\beta_2 x_i} e^{\beta_1+\beta_2 x_i}}{(1+e^{\beta_1+\beta_2 x_i})^2} \\
&= -(1-\gamma-\delta)\frac{e^{\beta_1+\beta_2 x_i}(1+e^{\beta_1+\beta_2 x_i} - e^{\beta_1+\beta_2 x_i})}{(1+e^{\beta_1+\beta_2 x_i})^2} \\
&= -(1-\gamma-\delta)\frac{e^{\beta_1+\beta_2 x_i}}{(1+e^{\beta_1+\beta_2 x_i})^2},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \beta_2} q_{x_i} &= -(1-\gamma-\delta)\frac{\partial}{\partial \beta_2}\pi_{x_i} \\
&= -(1-\gamma-\delta)\frac{x_i e^{\beta_1+\beta_2 x_i}(1+e^{\beta_1+\beta_2 x_i}) - x_i e^{\beta_1+\beta_2 x_i} e^{\beta_1+\beta_2 x_i}}{(1+e^{\beta_1+\beta_2 x_i})^2} \\
&= -(1-\gamma-\delta)\frac{x_i e^{\beta_1+\beta_2 x_i}(1+e^{\beta_1+\beta_2 x_i} - e^{\beta_1+\beta_2 x_i})}{(1+e^{\beta_1+\beta_2 x_i})^2} \\
&= -(1-\gamma-\delta)\frac{x_i e^{\beta_1+\beta_2 x_i}}{(1+e^{\beta_1+\beta_2 x_i})^2},
\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 q_{x_i}}{\partial \beta_1^2} &= -(1-\gamma-\delta)\frac{e^{\beta_1+\beta_2 x_i}(1+e^{\beta_1+\beta_2 x_i})^2 - 2(1+e^{\beta_1+\beta_2 x_i})(e^{\beta_1+\beta_2 x_i})^2}{(1+e^{\beta_1+\beta_2 x_i})^4} \\ &= -(1-\gamma-\delta)\frac{e^{\beta_1+\beta_2 x_i}(1+e^{\beta_1+\beta_2 x_i}-2e^{\beta_1+\beta_2 x_i})}{(1+e^{\beta_1+\beta_2 x_i})^3} \\ &= -(1-\gamma-\delta)\frac{e^{\beta_1+\beta_2 x_i}(1-e^{\beta_1+\beta_2 x_i})}{(1+e^{\beta_1+\beta_2 x_i})^3},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 q_{x_i}}{\partial \beta_2^2} &= -(1-\gamma-\delta)\frac{x_i^2 e^{\beta_1+\beta_2 x_i}(1+e^{\beta_1+\beta_2 x_i})^2}{(1+e^{\beta_1+\beta_2 x_i})^4} \\ &\quad +(1-\gamma-\delta)\frac{2(1+e^{\beta_1+\beta_2 x_i})x_i^2(e^{\beta_1+\beta_2 x_i})^2}{(1+e^{\beta_1+\beta_2 x_i})^4} \\ &= -(1-\gamma-\delta)\frac{x_i^2 e^{\beta_1+\beta_2 x_i}(1+e^{\beta_1+\beta_2 x_i}-2e^{\beta_1+\beta_2 x_i})}{(1+e^{\beta_1+\beta_2 x_i})^3} \\ &= -(1-\gamma-\delta)\frac{x_i^2 e^{\beta_1+\beta_2 x_i}(1-e^{\beta_1+\beta_2 x_i})}{(1+e^{\beta_1+\beta_2 x_i})^3},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 q_{x_i}}{\partial \beta_1 \beta_2} &= -(1-\gamma-\delta)\frac{x_i e^{\beta_1+\beta_2 x_i}(1+e^{\beta_1+\beta_2 x_i})^2}{(1+e^{\beta_1+\beta_2 x_i})^4} \\ &\quad +(1-\gamma-\delta)\frac{2(1+e^{\beta_1+\beta_2 x_i})x_i(e^{\beta_1+\beta_2 x_i})^2}{(1+e^{\beta_1+\beta_2 x_i})^4} \\ &= -(1-\gamma-\delta)\frac{x_i e^{\beta_1+\beta_2 x_i}(1+e^{\beta_1+\beta_2 x_i}-2e^{\beta_1+\beta_2 x_i})}{(1+e^{\beta_1+\beta_2 x_i})^3} \\ &= -(1-\gamma-\delta)\frac{x_i e^{\beta_1+\beta_2 x_i}(1-e^{\beta_1+\beta_2 x_i})}{(1+e^{\beta_1+\beta_2 x_i})^3} = \frac{\partial^2 q_{x_i}}{\partial \beta_2 \beta_1}.\end{aligned}$$

2.6.2 Il modello probit

Si consideri un modello probit con una variabile esplicativa scalare e intercetta zero. Si ha

$$\pi_{x_i} = \Phi(\beta x_i),$$

dove $\Phi(\cdot)$ indica la funzione di ripartizione di una distribuzione normale standard $N(0,1)$. Indicata con $\phi(\cdot)$ la densità di una normale standard e ricordando che

$$\frac{\partial}{\partial x}\Phi(x) = \phi(x)$$

e

$$\frac{\partial}{\partial x}\phi(x) = -x\phi(x),$$

si ottiene

$$\begin{aligned}\frac{\partial}{\partial\beta}q_{x_i} &= -(1-\gamma-\delta)\frac{\partial}{\partial\beta}\pi_{x_i}\\ &= -(1-\gamma-\delta)\frac{\partial}{\partial\beta}\Phi(\beta x_i)\\ &= -(1-\gamma-\delta)\phi(\beta x_i)x_i\end{aligned}$$

e

$$\begin{aligned}\frac{\partial^2 q_{x_i}}{\partial\beta^2} &= -(1-\gamma-\delta)x_i(-\beta x_i)\phi(\beta x_i)x_i\\ &= (1-\gamma-\delta)x_i^3\beta\phi(\beta x_i).\end{aligned}$$

Introducendo un parametro di intercetta, si ha $\beta = (\beta_1, \beta_2)$ e

$$\pi_{x_i} = \Phi(\beta_1 + \beta_2 x_i).$$

Si ha quindi

$$\begin{aligned}\frac{\partial}{\partial\beta_1}q_{x_i} &= -(1-\gamma-\delta)\frac{\partial}{\partial\beta_1}\pi_{x_i}\\ &= -(1-\gamma-\delta)\frac{\partial}{\partial\beta_1}\Phi(\beta_1+\beta_2 x_i)\\ &= -(1-\gamma-\delta)\phi(\beta_1+\beta_2 x_i),\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\beta_2}q_{x_i} &= -(1-\gamma-\delta)\frac{\partial}{\partial\beta_2}\pi_{x_i}\\ &= -(1-\gamma-\delta)\frac{\partial}{\partial\beta_2}\Phi(\beta_1+\beta_2 x_i)\\ &= -(1-\gamma-\delta)\phi(\beta_1+\beta_2 x_i)x_i,\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 q_{x_i}}{\partial\beta_1^2} &= -(1-\gamma-\delta)[-\phi(\beta_1+\beta_2 x_i)]\\ &= (1-\gamma-\delta)\phi(\beta_1+\beta_2 x_i),\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 q_{x_i}}{\partial \beta_2^2} &= -(1-\gamma-\delta)[-x_i\phi(\beta_1+\beta_2 x_i)] \\ &= (1-\gamma-\delta)x_i\phi(\beta_1+\beta_2 x_i),\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 q_{x_i}}{\partial \beta_1 \beta_2} &= -(1-\gamma-\delta)[-x_i\phi(\beta_1+\beta_2 x_i)] \\ &= (1-\gamma-\delta)x_i\phi(\beta_1+\beta_2 x_i) = \frac{\partial^2 q_{x_i}}{\partial \beta_2 \beta_1}.\end{aligned}$$

2.6.3 Il modello log-log complementare

Si consideri un modello log-log complementare con una variabile esplicativa scalare e intercetta zero. Si ha

$$\pi_{x_i} = 1 - e^{-e^{\beta x_i}}.$$

Risulta quindi

$$\begin{aligned}\frac{\partial}{\partial \beta} q_{x_i} &= -(1-\gamma-\delta)\frac{\partial}{\partial \beta}\pi_{x_i} \\ &= -(1-\gamma-\delta)\left[-e^{-e^{\beta x_i}}(-e^{\beta x_i})x_i\right] \\ &= -(1-\gamma-\delta)x_i e^{\beta x_i - e^{\beta x_i}}\end{aligned}$$

e

$$\begin{aligned}\frac{\partial^2 q_{x_i}}{\partial \beta^2} &= -(1-\gamma-\delta)x_i\frac{\partial}{\partial \beta}\left[e^{\beta x_i - e^{\beta x_i}}\right] \\ &= -(1-\gamma-\delta)x_i\left[e^{\beta x_i - e^{\beta x_i}}(x_i - x_i e^{\beta x_i})\right] \\ &= -(1-\gamma-\delta)x_i\left[x_i e^{\beta x_i - e^{\beta x_i}} - x_i e^{\beta x_i - e^{\beta x_i}} e^{\beta x_i}\right] \\ &= -(1-\gamma-\delta)x_i\left[x_i e^{\beta x_i - e^{\beta x_i}}(1 - e^{\beta x_i})\right] \\ &= -(1-\gamma-\delta)x_i^2 e^{\beta x_i - e^{\beta x_i}}(1 - e^{\beta x_i}).\end{aligned}$$

Introducendo un parametro di intercetta, si ha $\beta = (\beta_1, \beta_2)$ e

$$\pi_{x_i} = 1 - e^{-e^{\beta_1+\beta_2 x_i}}.$$

Si ha quindi

$$\begin{aligned}\frac{\partial}{\partial\beta_1}q_{x_i} &= -(1-\gamma-\delta)\frac{\partial}{\partial\beta_1}\pi_{x_i}\\ &= -(1-\gamma-\delta)\left(-e^{-e^{\beta_1+\beta_2 x_i}}\right)\left(-e^{\beta_1+\beta_2 x_i}\right)\\ &= -(1-\gamma-\delta)e^{-e^{\beta_1+\beta_2 x_i}}e^{\beta_1+\beta_2 x_i}\\ &= -(1-\gamma-\delta)e^{\beta_1+\beta_2 x_i-e^{\beta_1+\beta_2 x_i}},\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\beta_2}q_{x_i} &= -(1-\gamma-\delta)\frac{\partial}{\partial\beta_2}\pi_{x_i}\\ &= -(1-\gamma-\delta)\left(-e^{-e^{\beta_1+\beta_2 x_i}}\right)\left(-e^{\beta_1+\beta_2 x_i}x_i\right)\\ &= -(1-\gamma-\delta)x_i e^{-e^{\beta_1+\beta_2 x_i}}e^{\beta_1+\beta_2 x_i}\\ &= -(1-\gamma-\delta)x_i e^{\beta_1+\beta_2 x_i-e^{\beta_1+\beta_2 x_i}},\end{aligned}$$

$$\frac{\partial^2 q_{x_i}}{\partial\beta_1^2} = -(1-\gamma-\delta)e^{\beta_1+\beta_2 x_i-e^{\beta_1+\beta_2 x_i}}(1-e^{\beta_1+\beta_2 x_i}),$$

$$\frac{\partial^2 q_{x_i}}{\partial\beta_2^2} = -(1-\gamma-\delta)x_i e^{\beta_1+\beta_2 x_i-e^{\beta_1+\beta_2 x_i}}(x_i-x_i e^{\beta_1+\beta_2 x_i}),$$

$$\begin{aligned}\frac{\partial^2 q_{x_i}}{\partial\beta_1\beta_2} &= -(1-\gamma-\delta)x_i\left(e^{\beta_1+\beta_2 x_i-e^{\beta_1+\beta_2 x_i}}(1-e^{\beta_1+\beta_2 x_i})(x_i-x_i e^{\beta_1+\beta_2 x_i})\right)\\ &\quad -(1-\gamma-\delta)x_i\left(e^{\beta_1+\beta_2 x_i-e^{\beta_1+\beta_2 x_i}}(-x_i e^{\beta_1+\beta_2 x_i})\right) = \frac{\partial^2 q_{x_i}}{\partial\beta_2\beta_1}.\end{aligned}$$

2.7 Riferimenti bibliografici

Per le nozioni di base sulla classificazione scorretta, si è fatto riferimento a Küchenhoff (2007).
Per il paragrafo relativo ai modelli lineari generalizzati si è fatto riferimento ad Azzalini (2001, §§6.1, 6.2.1, 6.2.2) e a Nelder & Wedderburn (1972).
La sezioni relative alla specificazione della forma generale del modello, presentata nel paragrafo 2.4, ed ai casi particolari del modello, descritti nel paragrafo 2.6, sono state tratte da McInturff et al. (2004).

Per la correzione della distorsione con il metodo di Firth si è fatto nuovamente riferimento a Firth (1993). I calcoli svolti nei paragrafi 2.4, 2.5 e 2.6 sono originali.

Capitolo 3

Studi di simulazione

La simulazione, o la generazione di dati artificiali attraverso il computer, può essere usata per molti scopi: valutare la variabilità che ci si può attendere in un determinato modello, verificare l'adeguatezza di una determinata approssimazione teorica, controllare la sensitività dei risultati rispetto alle assunzioni, avere indicazioni di ricerca e fornire soluzioni numeriche quando non sono disponibili soluzioni analitiche. La simulazione usa dati generati casualmente per stimare o imitare i risultati di un processo casuale. I dati sono in realtà generati da un algoritmo deterministico, quindi i risultati sono (in linea di principio) prevedibili (Sartori, 2013).
In questo capitolo si esamina la teoria discussa fino ad ora attraverso degli studi di simulazione, che riguardano, nello specifico, la regressione logistica e la regressione probit.

3.1 Presentazione

Per valutare la validità del metodo di Firth (1993) nella riduzione della distorsione dello stimatore di massima verosimiglianza, sono stati condotti degli studi di simulazione basati su 5000 campioni casuali generati dalla variabile risposta $(Y_i \mid x_i) \sim$ Bernoulli(q_i) del modello di McInturff et al. (2004), presentato nel paragrafo 2.4, con $n = 100, 200, 500, 1000$, conside-

rando $\gamma = 0.90$ e $\delta = 0.80$ come parametri fissati, ed assegnando $\beta = 1$ e $\beta = (\beta_1, \beta_2) = (1, 1)$ come veri valori dei parametri da stimare, rispettivamente nel caso monoparametrico e nel caso biparametrico. Tali valori sono stati scelti facendo riferimento a Neuhaus (1999). In particolare, i valori per γ e δ rappresentano un caso estremo in cui le probabilità d'errore sono abbastanza elevate (0.10 e 0.20). Per ogni campione sono state calcolate le stime di massima verosimiglianza del parametro β, considerato dapprima come parametro scalare e poi come parametro bidimensionale. Successivamente è stata stimata la distorsione, $E(\hat{\beta}) - \beta$, sulla base dei 5000 campioni e si è studiato il comportamento di tale distorsione al crescere della numerosità campionaria n. Si sono inoltre riportati gli *standard error* stimati nelle simulazioni e le coperture stimate degli intervalli di confidenza alla Wald con livelli nominali 0.90, 0.95, 0.99, e 0.999.

3.2 Risultati

3.2.1 Regressione logistica

Nel primo studio di simulazione i campioni sono stati generati partendo da un modello logistico con β scalare, ossia considerando

$$\pi_{x_i} = \frac{e^{\beta x_i}}{1 + e^{\beta x_i}},$$

mentre nel secondo studio di simulazione si è considerato un parametro bidimensionale $\beta = (\beta_1, \beta_2)$, quindi si è considerato

$$\pi_{x_i} = \frac{e^{\beta_1 + \beta_2 x_i}}{1 + e^{\beta_1 + \beta_2 x_i}}.$$

Gli esiti del primo studio di simulazione sono riportati nelle quattro tabelle seguenti. In particolare, Le Tabelle 3.1 e 3.2 riportano le stime della distorsione e dello *standard error* dello stimatore di β. Come ci si attende, la distorsione dello stimatore di β decresce verso lo zero al crescere di n,

	Distorsione $\hat{\beta}$	Std Error $\hat{\beta}$
$n = 100$	0.1713	4.8354
$n = 200$	0.0447	0.3354
$n = 500$	0.0183	0.1990
$n = 1000$	0.0076	0.1365

Tabella 3.1: *Distorsione e standard error di $\hat{\beta}$ ottenuti nel modello logit a partire dalla funzione di punteggio non modificata nel caso in cui β è un parametro scalare.*

	Distorsione β^*	Std Error β^*
$n = 100$	0.0121	0.6684
$n = 200$	0.0081	0.3202
$n = 500$	0.0043	0.1955
$n = 1000$	0.0007	0.1353

Tabella 3.2: *Distorsione e standard error di β^* ottenuti nel modello logit a partire dalla funzione di punteggio modificata tramite il metodo di Firth (1993) nel caso in cui β è un parametro scalare.*

sia se stimata a partire dalla funzione di punteggio non modificata, sia se stimata a partire dalla funzione di punteggio modificata tramite il metodo di Firth (1993). Inoltre, la stima della distorsione risulta molto inferiore se calcolata a partire dalla *score* modificata, e ciò significa che il metodo di Firth (1993) ha successo nella rimozione del termine dominante della distorsione dello stimatore di massima verosimiglianza. Per quanto riguarda gli *standard error* dello stimatore, essi risultano essere più piccoli, e quindi migliori, se calcolati a partire dalla funzione di punteggio modificata. Lo *standard error* di $\hat{\beta}$ stimato a partire dalla *score* originale per $n = 100$ presenta un valore molto elevato (4.8354). Ciò è dovuto al fatto che, per $n = 100$, alcuni campioni risultano essere problematici, probabilmente per

la non esistenza finita della soluzione dell'equazione (l'algoritmo utilizzato da `R` nel pacchetto `nleqslv` non ha convergenza), ma lo stesso *standard error* calcolato a partire dalla funzione di punteggio modificata (0.6684) è basso, confermando ulteriormente la validità del metodo di Firth (1993).
Le Tabelle 3.3 e 3.4 seguenti riportano invece le coperture stimate degli intervalli di confidenza alla Wald di livelli nominali 0.90, 0.95, 0.99, e 0.999. Ci si aspetta che, ad esempio, a livello 0.90 il 90% dei valori osservati della statistica di Wald siano inferiori al quantile di un Chi-quadrato con un numero di gradi di libertà pari al numero di parametri da stimare (in questo caso uno). Dalle due tabelle sottostanti si può notare che sia le coperture stimate a partire dalla funzione di punteggio non modificata, sia quelle stimate a partire dalla funzione di punteggio modificata, si avvicinano ai livelli nominali. Tuttavia, le seconde sono leggermente migliori.

Livello di confidenza	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
0.90	0.9206	0.9170	0.9050	0.9026
0.95	0.9592	0.9616	0.9514	0.9498
0.99	0.9894	0.9862	0.9904	0.9906
0.999	0.9984	0.9976	0.9980	0.9990

Tabella 3.3: *Coperture stimate degli intervalli di confidenza alla Wald ottenute nel modello logit a partire dalla funzione di punteggio non modificata nel caso in cui β è un parametro scalare.*

Gli esiti del secondo studio di simulazione sono riportati nelle quattro tabelle seguenti. In particolare, Le Tabelle 3.5 e 3.6 riportano le stime della distorsione e degli *standard error* dello stimatore di $\beta = (\beta_1, \beta_2)$. Come nel caso scalare, anche nel caso in cui β è un parametro bidimensionale la distorsione decresce verso lo zero all'aumentare della numerosità campionaria. In particolare, sia la stima della distorsione dello stimatore di β_1, sia quella dello stimatore di β_2 risultano essere molto più piccole se calcolate a partire dalla funzione di punteggio modificata, confermando la

Livello di confidenza	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
0.90	0.9050	0.9100	0.9058	0.9022
0.95	0.9496	0.9544	0.9498	0.9500
0.99	0.9874	0.9844	0.9876	0.9902
0.999	0.9982	0.9962	0.9978	0.9986

Tabella 3.4: *Coperture stimate degli intervalli di confidenza alla Wald ottenute nel modello logit a partire dalla funzione di punteggio modificata tramite il metodo di Firth (1993) nel caso in cui β è un parametro scalare.*

	Distorsione $\hat{\beta}_1$	Std Error $\hat{\beta}_1$	Distorsione $\hat{\beta}_2$	Std Error $\hat{\beta}_2$
$n = 100$	0.0995	1.3802	0.1610	1.9130
$n = 200$	0.0389	0.2980	0.0522	0.3859
$n = 500$	0.0164	0.1787	0.0190	0.2162
$n = 1000$	0.0052	0.1220	0.0096	0.1463

Tabella 3.5: *Distorsione e standard error di $\hat{\beta}_1$ e $\hat{\beta}_2$ ottenuti nel modello logit a partire dalla funzione di punteggio non modificata nel caso in cui β è un parametro bidimensionale.*

validità del metodo di Firth (1993). Dalle stesse tabelle si nota che anche gli *standard error* risultano inferiori se stimati a partire dalla *score* modificata. Tuttavia, il metodo di Firth (1993) si focalizza sulla riduzione della distorsione, ed è per questo motivo che gli *standard error* risultano piccoli, e quindi adeguati, anche se calcolati a partire dalla funzione di punteggio non modificata.

Come si riscontra dalle Tabelle 3.7 e 3.8 riportate di seguito, le coperture degli intervalli di confidenza alla Wald stimate a partire dalla *score* non modificata risultano inferiori rispetto ai livelli nominali, soprattutto in corrispondenza di $n = 100$ e $n = 200$, mentre quelle stimate a partire dalla funzione di punteggio modificata si avvicinano molto ai livelli nominali.

	Distorsione β_1^*	Std Error β_1^*	Distorsione β_2^*	Std Error β_2^*
$n = 100$	-0.0015	0.4135	-0.0017	0.5621
$n = 200$	0.0001	0.2801	-0.0031	0.3596
$n = 500$	0.0023	0.1749	-0.0021	0.2104
$n = 1000$	0.0015	0.1208	-0.0003	0.1444

Tabella 3.6: *Distorsione e standard error di β_1^* e β_2^* ottenuti nel modello logit a partire dalla funzione di punteggio modificata tramite il metodo di Firth (1993) nel caso in cui β è un parametro bidimensionale.*

Ciò è dovuto principalmente alla riduzione della distorsione degli stimatori. C'è comunque da tener presente che stime non distorte non assicurano che i test siano accurati.

Liv.	$n = 100$		$n = 200$		$n = 500$		$n = 1000$	
conf.	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
0.90	0.8725	0.8665	0.8896	0.8902	0.8916	0.8962	0.9056	0.8980
0.95	0.9122	0.8953	0.9312	0.9210	0.9414	0.9454	0.9526	0.9436
0.99	0.9513	0.9368	0.9706	0.9578	0.9836	0.9776	0.9888	0.9856
0.999	0.9691	0.9603	0.9882	0.9800	0.9944	0.9922	0.9974	0.9974

Tabella 3.7: *Coperture stimate degli intervalli di confidenza alla Wald ottenute nel modello logit a partire dalla funzione di punteggio non modificata nel caso in cui β è un parametro bidimensionale.*

Per effettuare un'analisi più approfondita, si sono ripetuti gli stessi studi di simulazione con $n = 20$ e con $n = 50$. Con queste basse numerosità campionarie, i casi di possibile non convergenza si verificano più frequentemente, e talvolta la matrice di informazione attesa risulta non invertibile in R. Comunque, in accordo con le aspettative, più è piccola la numerosità campionaria, maggiori risultano le stime della distorsione e degli *standard error* dello stimatore. Tuttavia, le stime calcolate a partire dalla funzione

Liv. conf.	$n = 100$		$n = 200$		$n = 500$		$n = 1000$	
	β_1^*	β_2^*	β_1^*	β_2^*	β_1^*	β_2^*	β_1^*	β_2^*
0.90	0.9203	0.9227	0.9138	0.9178	0.8988	0.9152	0.9120	0.9034
0.95	0.9498	0.9415	0.9516	0.9434	0.9492	0.9582	0.9552	0.9496
0.99	0.9731	0.9642	0.9812	0.9706	0.9874	0.9844	0.9898	0.9892
0.999	0.9875	0.9800	0.9924	0.9902	0.9956	0.9944	0.9982	0.9974

Tabella 3.8: *Coperture stimate degli intervalli di confidenza alla Wald ottenute nel modello logit a partire dalla funzione di punteggio modificata tramite il metodo di Firth (1993) nel caso in cui β è un parametro bidimensionale.*

di punteggio modificata tramite il metodo di Firth (1993) risultano molto migliori rispetto alle stesse stime calcolate a partire dalla *score* non modificata.
Si può dunque concludere che, anche considerando delle numerosità campionarie poco elevate, il metodo proposto da Firth (1993) ha successo sia nel ridurre la distorsione dello stimatore di massima verosimiglianza, sia nel fornire, in talune situazioni problematiche, una funzione di verosimiglianza con massimo finito.

3.2.2 Regressione probit

Nel terzo studio di simulazione i campioni sono stati generati partendo da un modello probit con β scalare, ossia considerando

$$\pi_{x_i} = \Phi(\beta x_i),$$

mentre nel quarto studio di simulazione si è considerato un parametro bidimensionale $\beta = (\beta_1, \beta_2)$, quindi si è considerato

$$\pi_{x_i} = \Phi(\beta_1 + \beta_2 x_i).$$

Gli esiti del terzo studio di simulazione sono riportati nelle quattro tabelle

seguenti. In particolare, Le Tabelle 3.9 e 3.10 riportano le stime della distorsione e dello *standard error* dello stimatore di β. Osservando queste tabelle

	Distorsione $\hat{\beta}$	Std Error $\hat{\beta}$
$n = 100$	0.0733	17.9332
$n = 200$	0.0423	0.3216
$n = 500$	0.0197	0.1695
$n = 1000$	0.0076	0.1163

Tabella 3.9: *Distorsione e standard error di $\hat{\beta}$ ottenuti nel modello probit a partire dalla funzione di punteggio non modificata nel caso in cui β è un parametro scalare.*

	Distorsione β^*	Std Error β^*
$n = 100$	-0.0612	4.0190
$n = 200$	0.0026	0.2893
$n = 500$	0.0042	0.1655
$n = 1000$	0.0000	0.1150

Tabella 3.10: *Distorsione e standard error di β^* ottenuti nel modello probit a partire dalla funzione di punteggio modificata tramite il metodo di Firth (1993) nel caso in cui β è un parametro scalare.*

si nota che la distorsione dello stimatore, stimata sia a partire dalla funzione di punteggio non modificata, sia a partire dalla funzione di punteggio modificata, decresce verso lo zero all'aumentare della numerosità campionaria. Inoltre, la distorsione stimata a partire dalla *score* modificata risulta nettamente inferiore a quella stimata a partire dalla *score* originale, confermando, anche nel caso della regressione probit, la validità del metodo di Firth (1993). Si nota che, in corrispondenza di $n = 100$, lo *standard error* è molto elevato, sia se stimato a partire dalla funzione di punteggio non modificata (17.9332), sia se stimato a partire dalla funzione di punteggio

modificata (4.0190). Ciò è dovuto al fatto che alcune stime del parametro presentano dei valori piuttosto elevati, che a loro volta generano degli *standard error* molto elevati. Probabilmente ciò è dovuto alla non convergenza dell'algoritmo usato da R che, seguendo un criterio di arresto, si stoppa dopo un certo numero di iterazioni, anche se in realtà le stime non esistono finite. Sono quindi stati individuati i campioni problematici (5 nel caso in cui β è scalare) e sono stati eliminati dallo studio di simulazione, ottenendo due nuove stime della distorsione dello stimatore, pari a 0.0904 e a -0.0131 rispettivamente se calcolate a partire dalla funzione di punteggio non modificata e dalla funzione di punteggio modificata.
Le Tabelle 3.11 e 3.12 seguenti riportano invece le coperture stimate degli intervalli di confidenza alla Wald di livelli nominali 0.90, 0.95, 0.99, e 0.999. Da queste si può notare che sia le coperture stimate a partire dalla funzione di punteggio non modificata, sia quelle stimate a partire dalla funzione di punteggio modificata, si avvicinano ai livelli nominali, soprattutto per $n = 500$ e per $n = 1000$.

Livello di confidenza	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
0.90	0.9035	0.9114	0.9114	0.9046
0.95	0.9434	0.9520	0.9536	0.9504
0.99	0.9810	0.9824	0.9862	0.9890
0.999	0.9960	0.9946	0.9968	0.9980

Tabella 3.11: *Coperture stimate degli intervalli di confidenza alla Wald ottenute nel modello probit a partire dalla funzione di punteggio non modificata nel caso in cui β è un parametro scalare.*

Gli esiti del quarto studio di simulazione sono riportati nelle ultime quattro tabelle. In particolare, Le Tabelle 3.13 e 3.14 riportano le stime della distorsione e degli *standard error* dello stimatore di $\beta = (\beta_1, \beta_2)$. Come nel caso scalare, anche nel caso in cui β è un parametro bidimensionale la distorsione decresce verso lo zero all'aumentare della numerosità cam-

Livello di confidenza	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
0.90	0.8798	0.9018	0.9082	0.9006
0.95	0.9246	0.9410	0.9490	0.9476
0.99	0.9722	0.9766	0.9844	0.9878
0.999	0.9938	0.9940	0.9958	0.9978

Tabella 3.12: *Coperture stimate degli intervalli di confidenza alla Wald ottenute nel modello probit a partire dalla funzione di punteggio modificata tramite il metodo di Firth (1993) nel caso in cui β è un parametro scalare.*

	Distorsione $\hat{\beta}_1$	Std Error $\hat{\beta}_1$	Distorsione $\hat{\beta}_2$	Std Error $\hat{\beta}_2$
$n = 100$	-2.8930	111.5487	-6.3476	297.1848
$n = 200$	0.2192	11.8859	0.2600	13.3757
$n = 500$	0.0206	0.1627	0.0263	0.2004
$n = 1000$	0.0091	0.1107	0.0130	0.1339

Tabella 3.13: *Distorsione e standard error di $\hat{\beta}_1$ e $\hat{\beta}_2$ ottenuti nel modello probit a partire dalla funzione di punteggio non modificata nel caso in cui β è un parametro bidimensionale.*

pionaria. In particolare, sia la distorsione dello stimatore di β_1, sia quella dello stimatore di β_2 risultano essere molto più piccole se calcolate a partire dalla funzione di punteggio modificata. Partendo dalla *score* non modificata, in corrispondenza di $n = 100$, si ottengono delle distorsioni troppo elevate, sia per $\hat{\beta}_1$ (-2.8930), sia per $\hat{\beta}_2$ (-6.3476). Ciò è dovuto al fatto che, in corrispondenza di alcuni campioni, le stime dei due parametri non esistono finite. Tuttavia, l'algoritmo usato da R si ferma secondo un criterio di arresto, pur non convergendo, dopo un certo numero di iterazioni e genera quindi dei valori molto elevati di tali stime. Dunque si sono individuati i campioni problematici (69 nel caso biparametrico) e sono stati eliminati dallo studio di simulazione, ottenendo due nuove stime della

	Distorsione β_1^*	Std Error β_1^*	Distorsione β_2^*	Std Error β_2^*
$n = 100$	-0.0654	0.2362	-0.1916	0.2809
$n = 200$	-0.0199	0.2049	-0.0782	0.2421
$n = 500$	0.0038	0.1531	-0.0187	0.1790
$n = 1000$	0.0023	0.1086	-0.0061	0.1283

Tabella 3.14: *Distorsione e standard error di β_1^* e β_2^* ottenuti nel modello probit a partire dalla funzione di punteggio modificata tramite il metodo di Firth (1993) nel caso in cui β è un parametro bidimensionale.*

distorsione (0.1625 per $\hat{\beta}_1$ e 0.2214 per $\hat{\beta}_2$), molto più bassi dei precedenti. Per gli stessi campioni, le stime dei parametri ottenute a partire dalla funzione di punteggio modificata sono attendibili. Infatti, il vantaggio del metodo proposto da Firth (1993) va oltre la riduzione della distorsione e va individuato soprattutto nel fornire, in talune situazioni problematiche, una funzione di verosimiglianza con massimo finito. Tuttavia, per rendere possibile il confronto con le distorsioni ottenute a partire dalla *score* non modificata, questi campioni sono comunque stati eliminati dallo studio di simulazione, ottenendo come nuovi valori della distorsione -0.0681 per β_1^* e -0.1930 per β_2^*. Quindi, anche eliminando i campioni problematici, la distorsione che si ottiene a partire dalla funzione di punteggio modificata risulta inferiore a quella ottenuta a partire dalla *score* non modificata. Ciò conferma la validità del metodo di Firth (1993). Dalle stesse tabelle si nota che anche gli *standard error* risultano inferiori se stimati a partire dalla funzione di punteggio modificata.
Come si riscontra dalle Tabelle 3.15 e 3.16 riportate di seguito, le coperture degli intervalli di confidenza alla Wald stimate a partire dalla *score* non modificata risultano decisamente inferiori rispetto ai livelli nominali, soprattutto in corrispondenza di $n = 100$ e $n = 200$, mentre quelle stimate a partire dalla funzione di punteggio modificata, sempre in corrispondenza di $n = 100$ e $n = 200$, sono superiori ai livelli 0.90 e 0.95, ma, nei restanti

casi, si avvicinano ai livelli nominali.

Liv.	$n = 100$		$n = 200$		$n = 500$		$n = 1000$	
conf.	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
0.90	0.8278	0.7719	0.8748	0.8325	0.8921	0.8801	0.9050	0.8986
0.95	0.8648	0.8037	0.9126	0.8697	0.9341	0.9138	0.9488	0.9316
0.99	0.8989	0.8538	0.9529	0.9210	0.9709	0.9557	0.9818	0.9704
0.999	0.9255	0.8865	0.9745	0.9540	0.9890	0.9809	0.9940	0.9880

Tabella 3.15: *Coperture stimate degli intervalli di confidenza alla Wald ottenute nel modello probit a partire dalla funzione di punteggio non modificata nel caso in cui β è un parametro bidimensionale.*

Liv.	$n = 100$		$n = 200$		$n = 500$		$n = 1000$	
conf.	β_1^*	β_2^*	β_1^*	β_2^*	β_1^*	β_2^*	β_1^*	β_2^*
0.90	0.9845	0.9942	0.9528	0.9758	0.9160	0.9454	0.9138	0.9182
0.95	0.9966	0.9980	0.9768	0.9863	0.9482	0.9696	0.9570	0.9608
0.99	0.9989	0.9996	0.9964	0.9964	0.9826	0.9898	0.9852	0.9894
0.999	0.9993	1.0000	0.9996	0.9990	0.9942	0.9970	0.9950	0.9962

Tabella 3.16: *Coperture stimate degli intervalli di confidenza alla Wald ottenute nel modello probit a partire dalla funzione di punteggio modificata tramite il metodo di Firth (1993) nel caso in cui β è un parametro bidimensionale.*

Come per la regressione logistica, anche per la regressione probit si sono ripetuti gli stessi studi di simulazione con $n = 20$ e con $n = 50$. I risultati sono analoghi a quelli ottenuti per il modello logit: con basse numerosità campionarie, i casi di possibile non convergenza si verificano più frequentemente, e talvolta la matrice di informazione attesa risulta non invertibile in R. Comunque, in accordo con le aspettative, più è piccola la numerosità campionaria, maggiori risultano le stime della distorsione e degli *standard*

error dello stimatore. Tuttavia, le stime calcolate a partire dalla funzione di punteggio modificata tramite il metodo di Firth (1993) risultano molto migliori rispetto alle stesse stime calcolate a partire dalla *score* non modificata.
Si può dunque concludere che, anche considerando delle numerosità campionarie poco elevate, il metodo proposto da Firth (1993) ha successo sia nel ridurre la distorsione dello stimatore di massima verosimiglianza, sia nel fornire, in talune situazioni problematiche, una funzione di verosimiglianza con massimo finito.

3.3 Riferimenti bibliografici

Per la sezione relativa alla breve presentazione e allo scopo d'uso delle simulazioni si è fatto riferimento a Sartori (2013, pagg. 6-8). Il resto del capitolo è originale.

Conclusioni

In questa tesi si è voluta evidenziare la validità del metodo proposto da Firth (1993) per ridurre la distorsione dello stimatore di massima verosimiglianza in modelli per la classificazione scorretta di dati binari. Il metodo di Firth (1993) consiste essenzialmente in una modificazione del meccanismo che produce la stima di massima verosimiglianza, ossia dell'equazione di verosimiglianza basata sulla funzione di punteggio, piuttosto che della stima stessa. Il vantaggio di tale approccio va oltre la semplice riduzione della distorsione dello stimatore di massima verosimiglianza, e va individuato soprattutto nel fornire, in talune situazioni problematiche, una funzione di verosimiglianza con massimo finito.
Per errata classificazione si intende l'assegnazione dei soggetti di studio alla categoria sbagliata di una variabile categoriale. Dunque, è considerabile come un *errore di classificazione*.
Partendo dal modello proposto da McInturff et al. (2004) si sono effettuati degli studi di simulazione tramite l'utilizzo dell'ambiente R. Dapprima si è considerata la funzione di legame di tipo logit, distinguendo il caso con un solo parametro dal caso con due parametri, e successivamente si è ripetuto lo stesso tipo di studio considerando la funzione di legame di tipo probit. In entrambi i casi, si è stimata la distorsione dello stimatore di massima verosimiglianza sia a partire dalla funzione di punteggio non modificata, sia a partire dalla funzione di punteggio modificata tramite il metodo di Firth (1993). Come ci si attende, la distorsione dello stimatore decresce verso lo zero al crescere della numerosità campionaria, sia se

calcolata a partire dalla *score* non modificata, sia se calcolata a partire dalla *score* modificata. Inoltre, la distorsione risulta molto inferiore se calcolata a partire dalla funzione di punteggio modificata, e ciò significa che il metodo di Firth (1993) ha successo nella rimozione del termine dominante della distorsione dello stimatore di massima verosimiglianza. Nel caso della regressione probit, utilizzando la *score* non modificata, alcuni campioni sono risultati problematici, non esistendo finite le stime dei parametri per tali campioni. Quindi si sono individuati tali campioni e si sono esclusi dallo studio di simulazione, ottenendo, senza di essi, dei valori più attendibili della distorsione. Per gli stessi campioni, si è visto che le stime ottenute a partire dalla funzione di punteggio modificata erano attendibili, ma si sono comunque esclusi per rendere possibile il confronto con la distorsione ottenuta a partire dalla *score* non modificata. Di nuovo, la distorsione che si ottiene dalla funzione di punteggio modificata è risultata nettamente inferiore a quella che si ottiene dalla funzione di punteggio non modificata. Si può dunque ritenere che il metodo proposto da Firth (1993) costituisca un valido strumento per ridurre la distorsione dello stimatore di massima verosimiglianza. L'utilizzo delle formule relative alla regressione log-log complementare presentate nel corso di questa tesi per effettuare un analogo studio di simulazione potrebbe essere oggetto di analisi successive e ulteriore conferma della validità del metodo di Firth (1993).

Appendice

Di seguito, si riportano i comandi `R` utilizzati per effettuare gli studi di simulazione del Capitolo 3. Le variabili `eta` e `theta` corrispondono rispettivamente a γ e δ utilizzati nella tesi.

```
# CORREZIONE CON IL METODO DI FIRTH

# LOGIT CON UN SOLO PARAMETRO:

n <- 100
eta0 <- 0.90
theta0 <- 0.80
set.seed(123)
x0 <- rnorm(n)
beta0 <- 1

csi0 <- beta0*x0

pi0logit <- exp(csi0)/(1+exp(csi0))
q0 <- 1-theta0-(1-eta0-theta0)*pi0logit

y0 <- rbinom(n,1,q0)

Ub <- function(y, beta, eta0, theta0, x) {
csi <- beta*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
a <- V1/qx
b <- V1/(1-qx)
```

```
sum((a+b)*y)-sum(b)
}

Ubb <- function(y, beta, eta0, theta0, x) {
csi <- beta*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
V2 <- -(1-eta0-theta0)*x^2*exp(csi)*(1-exp(csi))/(1+exp(csi))^3
c <- (V2*qx-V1^2)/qx^2
d <- (V2*(1-qx)+V1^2)/(1-qx)^2
sum((c+d)*y)-sum(d)
}

Kb_b <- function(n, beta, eta0, theta0, x) {
csi <- beta*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
a <- V1/qx
b <- V1/(1-qx)
mat <- (a+b)^2*qx+b^2-2*(a+b)*b*qx
(1/n)*sum(mat)
}

Kb_b.inv <- 1/Kb_b(n,beta0,eta0,theta0,x0)

Kb_b_b <- function(n, beta, eta0, theta0, x) {
csi <- beta*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
a <- V1/qx
b <- V1/(1-qx)
mat <- (a+b)^3*qx-b^3-3*(a+b)^2*b*qx+3*(a+b)*b^2*qx
(1/n)*sum(mat)
}
```

```
Kb_bb <- function (n, beta, eta0, theta0, x) {
csi <- beta*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
a <- V1/qx
b <- V1/(1-qx)
V2 <- -(1-eta0-theta0)*x^2*exp(csi)*(1-exp(csi))/(1+exp(csi))^3
c <- (V2*qx-V1^2)/qx^2
d <- (V2*(1-qx)+V1^2)/(1-qx)^2
mat <- (a+b)*(c+d)*qx-(a+b)*d*qx-(c+d)*b*qx+b*d
(1/n)*sum(mat)
}

Ab <- function(n,beta0,eta0,theta0,x0,y0,Kb_b.inv) {
-Ubb(y0,beta0,eta0,theta0,x0)/(2*n)*Kb_b.inv^2*
(Kb_b_b(n,beta0,eta0,theta0,x0)+Kb_bb(n,beta0,eta0,theta0,x0))
}

library(nleqslv)

# Con la funzione di punteggio modificata (metodo di Firth):

U.star <- function(beta0, eta0, theta0, x0, y0) {
n = length(y0)

Kb_b.inv <- 1/Kb_b(n,beta0,eta0,theta0,x0)

y = numeric(1)

ub = Ub(y0, beta0, eta0, theta0, x0)

ab = Ab(n,beta0,eta0,theta0,x0,y0,Kb_b.inv)

y = ub + ab

y
}
```

```
nleqslv(0,U.star,eta0=eta0,theta0=theta0,x0=x0,y0=y0,
control=list(btol=.01))

stat1par<-function(R,n,beta0,eta0,theta0,conf.level=c(0.90,
0.95,0.99,0.999)) {
set.seed(123)
x0 <- rnorm(n)
std.error.beta<-wald<-rep(NA,R)
mat.est <- matrix(0,1,R)
for(i in 1:R) {
csi0 <- beta0*x0
pi0logit <- exp(csi0)/(1+exp(csi0))
q0 <- 1-theta0-(1-eta0-theta0)*pi0logit
y0 <- rbinom(n,1,q0)
mat.est[,i] = nleqslv(0, U.star, eta0=eta0,
theta0=theta0, x0=x0, y0=y0, control=list(btol=.01))$x
std.error.beta[i] <- sqrt(-1/grad(U.star,mat.est[1,i],
eta0=eta0,theta0=theta0,y0=y0,x0=x0))
wald[i] <- (mat.est[1,i] - beta0)^2/std.error.beta[i]^2
}
wald.coverage <- sapply(conf.level, function(x0) mean(wald <
qchisq(x0, 1)))
list(Bias=apply(mat.est,1,mean)-beta0, Mat.est=mat.est,
Wald.coverage=wald.coverage,
std.error=apply(mat.est,1,function(x) sqrt(var(x))))
}

prova1par100 <- stat1par(5000,100,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999))
prova1par200 <- stat1par(5000,200,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999))
prova1par500 <- stat1par(5000,500,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999))
prova1par1000 <- stat1par(5000,1000,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999))

save(prova1par100,file="prova1par100.RData")
```

```
save(prova1par200,file="prova1par200.RData")
save(prova1par500,file="prova1par500.RData")
save(prova1par1000,file="prova1par1000.RData")

# Con la funzione di punteggio non modificata:

U.star.distorta <- function(beta0, eta0, theta0, x0, y0) {
n = length(y0)

y = numeric(1)

ub = Ub(y0, beta0, eta0, theta0, x0)

y = ub

y
}

nleqslv(0, U.star.distorta, eta0=eta0, theta0=theta0,
x0=x0, y0=y0, control=list(btol=.01))

stat1par.nomod<-function(R,n,beta0,eta0,theta0,conf.level=
c(0.90,0.95,0.99,0.999)) {
set.seed(123)
x0 <- rnorm(n)
std.error.beta<-wald<-rep(NA,R)
mat.est <- matrix(0,1,R)
for(i in 1:R) {
csi0 <- beta0*x0
pi0logit <- exp(csi0)/(1+exp(csi0))
q0 <- 1-theta0-(1-eta0-theta0)*pi0logit
y0 <- rbinom(n,1,q0)
mat.est[,i] = nleqslv(0, U.star.distorta, eta0=eta0,
theta0=theta0, x0=x0, y0=y0, control=list(btol=.01))$x
std.error.beta[i] <- sqrt(-1/Ubb(y0,mat.est[1,i],eta0,
theta0,x0))
wald[i] <- (mat.est[1,i] - beta0)^2/std.error.beta[i]^2
}
```

```
wald.coverage <- sapply(conf.level, function(x0) mean(wald <
qchisq(x0, 1)))
list(Bias=apply(mat.est,1,mean)-beta0, Mat.est=mat.est,
Wald.coverage=wald.coverage,
std.error=apply(mat.est,1,function(x) sqrt(var(x))))
}

prova2.1par100 <- stat1par.nomod(5000,100,beta0,eta0,
theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova2.1par200 <- stat1par.nomod(5000,200,beta0,eta0,
theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova2.1par500 <- stat1par.nomod(5000,500,beta0,eta0,
theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova2.1par1000 <- stat1par.nomod(5000,1000,beta0,eta0,
theta0,conf.level=c(0.90,0.95,0.99,0.999))

save(prova2.1par100,file="prova2.1par100.RData")
save(prova2.1par200,file="prova2.1par200.RData")
save(prova2.1par500,file="prova2.1par500.RData")
save(prova2.1par1000,file="prova2.1par1000.RData")


# LOGIT CON DUE PARAMETRI:

n <- 100
eta0 <- 0.90
theta0 <- 0.80
set.seed(123)
x0 <- rnorm(n)
beta1 <- 1
beta2 <- 1
beta0 <- c(beta1,beta2)

csi0 <- beta0[1]+beta0[2]*x0

pi0logit <- exp(csi0)/(1+exp(csi0))
q0 <- 1-theta0-(1-eta0-theta0)*pi0logit
```

```
y0 <- rbinom(n,1,q0)

U1 <- function(y, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*exp(csi)/(1+exp(csi))^2
a1 <- V1/qx
b1 <- V1/(1-qx)
sum((a1+b1)*y)-sum(b1)
}

U2 <- function(y, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V2 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
a2 <- V2/qx
b2 <- V2/(1-qx)
sum((a2+b2)*y)-sum(b2)
}

U11 <- function(y, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*exp(csi)/(1+exp(csi))^2
V3 <- -(1-eta0-theta0)*exp(csi)*(1-exp(csi))/(1+exp(csi))^3
c1 <- (V3*qx-V1^2)/qx^2
d1 <- (V3*(1-qx)+V1^2)/(1-qx)^2
sum((c1+d1)*y)-sum(d1)
}

U22 <- function(y, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V2 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
```

```
V4 <- -(1-eta0-theta0)*x^2*exp(csi)*(1-exp(csi))/(1+exp(csi))^3
c2 <- (V4*qx-V2^2)/qx^2
d2 <- (V4*(1-qx)+V2^2)/(1-qx)^2
sum((c2+d2)*y)-sum(d2)
}

U12 <- function(y, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*exp(csi)/(1+exp(csi))^2
V2 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
V5 <- -(1-eta0-theta0)*x*exp(csi)*(1-exp(csi))/(1+exp(csi))^3
c12 <- (V5*qx-V1*V2)/qx^2
d12 <- (V5*(1-qx)+V1*V2)/(1-qx)^2
sum((c12+d12)*y)-sum(d12)
}

K1_1 <- function(n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*exp(csi)/(1+exp(csi))^2
a1 <- V1/qx
b1 <- V1/(1-qx)
mat <- (a1+b1)^2*qx+b1^2-2*(a1+b1)*b1*qx
(1/n)*sum(mat)
}

K2_2 <- function(n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V2 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
a2 <- V2/qx
b2 <- V2/(1-qx)
mat <- (a2+b2)^2*qx+b2^2-2*(a2+b2)*b2*qx
(1/n)*sum(mat)
```

```
}

K1_2 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*exp(csi)/(1+exp(csi))^2
a1 <- V1/qx
b1 <- V1/(1-qx)
V2 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
a2 <- V2/qx
b2 <- V2/(1-qx)
mat <- (a1+b1)*(a2+b2)*qx-(a1+b1)*b2*qx-(a2+b2)*b1*qx+b1*b2
(1/n)*sum(mat)
}

K <- matrix(c(K1_1(n,beta0,eta0,theta0,x0),
K1_2(n,beta0,eta0,theta0,x0),K1_2(n,beta0,eta0,theta0,x0),
K2_2(n,beta0,eta0,theta0,x0)),2,2)

Kinv <- solve(K)

K1_1_1 <- function(n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*exp(csi)/(1+exp(csi))^2
a1 <- V1/qx
b1 <- V1/(1-qx)
mat <- (a1+b1)^3*qx-b1^3-3*(a1+b1)^2*b1*qx+3*(a1+b1)*b1^2*qx
(1/n)*sum(mat)
}

K2_2_2 <- function(n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V2 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
```

```
a2 <- V2/qx
b2 <- V2/(1-qx)
mat <- (a2+b2)^3*qx-b2^3-3*(a2+b2)^2*b2*qx+3*(a2+b2)*b2^2*qx
(1/n)*sum(mat)
}

K1_11 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*exp(csi)/(1+exp(csi))^2
a1 <- V1/qx
b1 <- V1/(1-qx)
V3 <- -(1-eta0-theta0)*exp(csi)*(1-exp(csi))/(1+exp(csi))^3
c1 <- (V3*qx-V1^2)/qx^2
d1 <- (V3*(1-qx)+V1^2)/(1-qx)^2
mat <- (a1+b1)*(c1+d1)*qx-(a1+b1)*d1*qx-(c1+d1)*b1*qx+b1*d1
(1/n)*sum(mat)
}

K1_1_2 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*exp(csi)/(1+exp(csi))^2
a1 <- V1/qx
b1 <- V1/(1-qx)
V2 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
a2 <- V2/qx
b2 <- V2/(1-qx)
mat <- (a1+b1)^2*(a2+b2)*qx-(a1+b1)^2*b2*qx+(a2+b2)*b1^2*qx-
b1^2*b2-2*(a1+b1)*b1*(a2+b2)*qx+2*(a1+b1)*b1*b2*qx
(1/n)*sum(mat)
}

K1_12 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
```

```
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*exp(csi)/(1+exp(csi))^2
a1 <- V1/qx
b1 <- V1/(1-qx)
V2 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
V5 <- -(1-eta0-theta0)*x*exp(csi)*(1-exp(csi))/(1+exp(csi))^3
c12 <- (V5*qx-V1*V2)/qx^2
d12 <- (V5*(1-qx)+V1*V2)/(1-qx)^2
mat <- (a1+b1)*(c12+d12)*qx-(a1+b1)*d12*qx-(c12+d12)*b1*qx+b1*d12
(1/n)*sum(mat)
}

K1_2_2 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*exp(csi)/(1+exp(csi))^2
a1 <- V1/qx
b1 <- V1/(1-qx)
V2 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
a2 <- V2/qx
b2 <- V2/(1-qx)
mat <- (a1+b1)*(a2+b2)^2*qx-(a2+b2)^2*b1*qx+(a1+b1)*b2^2*qx-
b1*b2^2-2*(a2+b2)*b2*(a1+b1)*qx+2*(a2+b2)*b1*b2*qx
(1/n)*sum(mat)
}

K1_22 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V1 <- -(1-eta0-theta0)*exp(csi)/(1+exp(csi))^2
a1 <- V1/qx
b1 <- V1/(1-qx)
V2 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
V4 <- -(1-eta0-theta0)*x^2*exp(csi)*(1-exp(csi))/(1+exp(csi))^3
c2 <- (V4*qx-V2^2)/qx^2
d2 <- (V4*(1-qx)+V2^2)/(1-qx)^2
```

```
mat <- (a1+b1)*(c2+d2)*qx-(a1+b1)*d2*qx-(c2+d2)*b1*qx+b1*d2
(1/n)*sum(mat)
}

K2_11 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V2 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
a2 <- V2/qx
b2 <- V2/(1-qx)
V1 <- -(1-eta0-theta0)*exp(csi)/(1+exp(csi))^2
V3 <- -(1-eta0-theta0)*exp(csi)*(1-exp(csi))/(1+exp(csi))^3
c1 <- (V3*qx-V1^2)/qx^2
d1 <- (V3*(1-qx)+V1^2)/(1-qx)^2
mat <- (a2+b2)*(c1+d1)*qx-(a2+b2)*d1*qx-(c1+d1)*b2*qx+b2*d1
(1/n)*sum(mat)
}

K2_12 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
V2 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
a2 <- V2/qx
b2 <- V2/(1-qx)
V1 <- -(1-eta0-theta0)*exp(csi)/(1+exp(csi))^2
V5 <- -(1-eta0-theta0)*x*exp(csi)*(1-exp(csi))/(1+exp(csi))^3
c12 <- (V5*qx-V1*V2)/qx^2
d12 <- (V5*(1-qx)+V1*V2)/(1-qx)^2
mat <- (a2+b2)*(c12+d12)*qx-(a2+b2)*d12*qx-(c12+d12)*b2*qx+b2*d12
(1/n)*sum(mat)
}

K2_22 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_logit <- exp(csi)/(1+exp(csi))
qx <- 1-theta0-(1-eta0-theta0)*pi_logit
```

```
V2 <- -(1-eta0-theta0)*x*exp(csi)/(1+exp(csi))^2
a2 <- V2/qx
b2 <- V2/(1-qx)
V4 <- -(1-eta0-theta0)*x^2*exp(csi)*(1-exp(csi))/(1+exp(csi))^3
c2 <- (V4*qx-V2^2)/qx^2
d2 <- (V4*(1-qx)+V2^2)/(1-qx)^2
mat <- (a2+b2)*(c2+d2)*qx-(a2+b2)*d2*qx-(c2+d2)*b2*qx+b2*d2
(1/n)*sum(mat)
}

A1 <- function(n,beta0,eta0,theta0,x0,y0,Kinv) {
-U11(y0,beta0,eta0,theta0,x0)/(2*n)*(Kinv[1,1]*Kinv[1,1]*
(K1_1_1(n,beta0,eta0,theta0,x0)+K1_11(n,beta0,eta0,theta0,x0))+
Kinv[1,1]*Kinv[1,2]*(K1_1_2(n,beta0,eta0,theta0,x0)+
K1_12(n,beta0,eta0,theta0,x0))+Kinv[1,1]*Kinv[1,2]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K1_12(n,beta0,eta0,theta0,x0))+
Kinv[1,1]*Kinv[2,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K1_22(n,beta0,eta0,theta0,x0))+Kinv[1,2]*Kinv[1,1]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K2_11(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[1,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K2_12(n,beta0,eta0,theta0,x0))+Kinv[1,2]*Kinv[1,2]*
(K1_2_2(n,beta0,eta0,theta0,x0)+K2_12(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[2,2]*(K2_2_2(n,beta0,eta0,theta0,x0)+
K2_22(n,beta0,eta0,theta0,x0)))-
U12(y0,beta0,eta0,theta0,x0)/(2*n)*(Kinv[1,2]*Kinv[1,1]*
(K1_1_1(n,beta0,eta0,theta0,x0)+K1_11(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[1,2]*(K1_1_2(n,beta0,eta0,theta0,x0)+
K1_12(n,beta0,eta0,theta0,x0))+Kinv[1,2]*Kinv[1,2]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K1_12(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[2,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K1_22(n,beta0,eta0,theta0,x0))+Kinv[2,2]*Kinv[1,1]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K2_11(n,beta0,eta0,theta0,x0))+
Kinv[2,2]*Kinv[1,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K2_12(n,beta0,eta0,theta0,x0))+Kinv[2,2]*Kinv[1,2]*
(K1_2_2(n,beta0,eta0,theta0,x0)+K2_12(n,beta0,eta0,theta0,x0))+
Kinv[2,2]*Kinv[2,2]*(K2_2_2(n,beta0,eta0,theta0,x0)+
K2_22(n,beta0,eta0,theta0,x0)))
}
```

```
A2 <- function(n,beta0,eta0,theta0,x0,y0,Kinv) {
-U12(y0,beta0,eta0,theta0,x0)/(2*n)*(Kinv[1,1]*Kinv[1,1]*
(K1_1_1(n,beta0,eta0,theta0,x0)+K1_11(n,beta0,eta0,theta0,x0))+
Kinv[1,1]*Kinv[1,2]*(K1_1_2(n,beta0,eta0,theta0,x0)+
K1_12(n,beta0,eta0,theta0,x0))+Kinv[1,1]*Kinv[1,2]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K1_12(n,beta0,eta0,theta0,x0))+
Kinv[1,1]*Kinv[2,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K1_22(n,beta0,eta0,theta0,x0))+Kinv[1,2]*Kinv[1,1]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K2_11(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[1,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K2_12(n,beta0,eta0,theta0,x0))+Kinv[1,2]*Kinv[1,2]*
(K1_2_2(n,beta0,eta0,theta0,x0)+K2_12(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[2,2]*(K2_2_2(n,beta0,eta0,theta0,x0)+
K2_22(n,beta0,eta0,theta0,x0)))-
U22(y0,beta0,eta0,theta0,x0)/(2*n)*(Kinv[1,2]*Kinv[1,1]*
(K1_1_1(n,beta0,eta0,theta0,x0)+K1_11(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[1,2]*(K1_1_2(n,beta0,eta0,theta0,x0)+
K1_12(n,beta0,eta0,theta0,x0))+Kinv[1,2]*Kinv[1,2]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K1_12(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[2,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K1_22(n,beta0,eta0,theta0,x0))+Kinv[2,2]*Kinv[1,1]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K2_11(n,beta0,eta0,theta0,x0))+
Kinv[2,2]*Kinv[1,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K2_12(n,beta0,eta0,theta0,x0))+Kinv[2,2]*Kinv[1,2]*
(K1_2_2(n,beta0,eta0,theta0,x0)+K2_12(n,beta0,eta0,theta0,x0))+
Kinv[2,2]*Kinv[2,2]*(K2_2_2(n,beta0,eta0,theta0,x0)+
K2_22(n,beta0,eta0,theta0,x0)))
}

library(nleqslv)

# Con la funzione di punteggio modificata (metodo di Firth):

u.star.beta <- function(beta0, eta0, theta0, x0, y0) {
n = length(y0)

K <- matrix(c(K1_1(n,beta0,eta0,theta0,x0),
```

```
K1_2(n,beta0,eta0,theta0,x0),K1_2(n,beta0,eta0,theta0,x0),
K2_2(n,beta0,eta0,theta0,x0)),2,2)
Kinv <- solve(K)

y = numeric(2)

u1 = U1(y0, beta0, eta0, theta0, x0)
u2 = U2(y0, beta0, eta0, theta0, x0)

a1 = A1(n,beta0,eta0,theta0,x0,y0,Kinv)
a2 = A2(n,beta0,eta0,theta0,x0,y0,Kinv)

y[1] = u1 + a1
y[2] = u2 + a2

y
}

nleqslv(c(0,0), u.star.beta, eta0=eta0, theta0=theta0,
x0=x0, y0=y0, control=list(btol=.01))

library(numDeriv)

# Matrice di informazione osservata inversa:

funct.genD <- function(beta0, eta0, theta0, x0, y0) {
-u.star.beta(c(beta0[1],beta0[2]), eta0, theta0, x0, y0)
}

info.oss.firth.inv <- function(beta0, eta0, theta0, x0, y0){
ris<- genD(funct.genD,c(1,1),eta0=eta0,theta0=theta0,x0=x0,
y0=y0)$D[,1:2]
solve(ris)
}

statistiche<-function(R,n,beta0,eta0,theta0,conf.level=c(0.90,
0.95,0.99,0.999)) {
set.seed(123)
```

```
x0 <- rnorm(n)
mat.est <- matrix(0,2,R)
std.error.beta1<-std.error.beta2<-wald1<-wald2<-rep(NA,R)
for(i in 1:R) {
csi0 <- beta0[1]+beta0[2]*x0
pi0logit <- exp(csi0)/(1+exp(csi0))
q0 <- 1-theta0-(1-eta0-theta0)*pi0logit
y0 <- rbinom(n,1,q0)
info <- sqrt(diag(info.oss.firth.inv(beta0,eta0,theta0,x0,y0)))
std.error.beta1[i] <- info[1]
std.error.beta2[i] <- info[2]
mat.est[,i] = nleqslv(c(0,0), u.star.beta, eta0=eta0,
theta0=theta0, x0=x0, y0=y0, control=list(btol=.01))$x
wald1[i] <- (mat.est[1,i] - beta0[1])^2/std.error.beta1[i]^2
wald2[i] <- (mat.est[2,i] - beta0[2])^2/std.error.beta2[i]^2
}
wald.coverage1 <- sapply(conf.level, function(x0) mean(wald1
[which(!is.na(wald1))] < qchisq(x0, 1)))
wald.coverage2 <- sapply(conf.level, function(x0) mean(wald2
[which(!is.na(wald2))] < qchisq(x0, 1)))
list(Bias=apply(mat.est,1,mean)-beta0, Mat.est=mat.est,
Wald.coverage1=wald.coverage1,
Wald.coverage2=wald.coverage2,
std.error=apply(mat.est,1,function(x) sqrt(var(x))))
}

prova100 <- statistiche(5000,100,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999))
prova200 <- statistiche(5000,200,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999))
prova500 <- statistiche(5000,500,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999))
prova1000 <- statistiche(5000,1000,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999))

save(prova100,file="prova100.RData")
save(prova200,file="prova200.RData")
save(prova500,file="prova500.RData")
```

```
save(prova1000,file="prova1000.RData")


# Con la funzione di punteggio non modificata:

u.star.beta.distorta <- function(beta0, eta0, theta0, x0, y0) {
n = length(y0)

y = numeric(2)

u1 = U1(y0, beta0, eta0, theta0, x0)
u2 = U2(y0, beta0, eta0, theta0, x0)

y[1] = u1
y[2] = u2

y
}

nleqslv(c(0,0), u.star.beta.distorta, eta0=eta0, theta0=theta0,
x0=x0, y0=y0, control=list(btol=.01))

# Matrice di informazione osservata inversa:

info.oss.inv <- function(beta0, eta0, theta0, x0, y0) {
n = length(y0)
solve(-matrix(c(U11(y0,beta0,eta0,theta0,x0),U12(y0,
beta0,eta0,theta0,x0),U12(y0,beta0,eta0,theta0,x0),
U22(y0,beta0,eta0,theta0,x0)),2,2))
}

statistiche.nomod<-function(R,n,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999)) {
set.seed(123)
x0 <- rnorm(n)
mat.est <- matrix(0,2,R)
std.error.beta1<-std.error.beta2<-wald1<-wald2<-rep(NA,R)
for(i in 1:R) {
```

```
csi0 <- beta0[1]+beta0[2]*x0
pi0logit <- exp(csi0)/(1+exp(csi0))
q0 <- 1-theta0-(1-eta0-theta0)*pi0logit
y0 <- rbinom(n,1,q0)
info <- sqrt(diag(info.oss.inv(beta0,eta0,theta0,x0,y0)))
std.error.beta1[i] <- info[1]
std.error.beta2[i] <- info[2]
mat.est[,i] = nleqslv(c(0,0), u.star.beta.distorta, eta0=eta0,
theta0=theta0, x0=x0, y0=y0, control=list(btol=.01))$x
wald1[i] <- (mat.est[1,i] - beta0[1])^2/std.error.beta1[i]^2
wald2[i] <- (mat.est[2,i] - beta0[2])^2/std.error.beta2[i]^2
}
wald.coverage1 <- sapply(conf.level, function(x0) mean(wald1
[which(!is.na(wald1))] < qchisq(x0, 1)))
wald.coverage2 <- sapply(conf.level, function(x0) mean(wald2
[which(!is.na(wald2))] < qchisq(x0, 1)))
list(Bias=apply(mat.est,1,mean)-beta0, Mat.est=mat.est,
Wald.coverage1=wald.coverage1,
Wald.coverage2=wald.coverage2,
std.error=apply(mat.est,1,function(x) sqrt(var(x))))
}

prova2_100 <- statistiche.nomod(5000,100,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999))
prova2_200 <- statistiche.nomod(5000,200,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999))
prova2_500 <- statistiche.nomod(5000,500,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999))
prova2_1000 <- statistiche.nomod(5000,1000,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999))

save(prova2_100,file="prova2_100.RData")
save(prova2_200,file="prova2_200.RData")
save(prova2_500,file="prova2_500.RData")
save(prova2_1000,file="prova2_1000.RData")


# PROBIT CON UN SOLO PARAMETRO:
```

```
n <- 100
eta0 <- 0.90
theta0 <- 0.80
set.seed(123)
x0 <- rnorm(n)
beta0 <- 1

csi0 <- beta0*x0

pi0probit <- pnorm(csi0)
q0 <- 1-theta0-(1-eta0-theta0)*pi0probit

y0 <- rbinom(n,1,q0)

Ub <- function(y, beta, eta0, theta0, x) {
csi <- beta*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*x*dnorm(csi)
a <- V1/qx
b <- V1/(1-qx)
sum((a+b)*y)-sum(b)
}

Ubb <- function(y, beta, eta0, theta0, x) {
csi <- beta*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*x*dnorm(csi)
V2 <- (1-eta0-theta0)*x^2*dnorm(csi)*csi
c <- (V2*qx-V1^2)/qx^2
d <- (V2*(1-qx)+V1^2)/(1-qx)^2
sum((c+d)*y)-sum(d)
}

Kb_b <- function(n, beta, eta0, theta0, x) {
csi <- beta*x
```

```
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*x*dnorm(csi)
a <- V1/qx
b <- V1/(1-qx)
mat <- (a+b)^2*qx+b^2-2*(a+b)*b*qx
(1/n)*sum(mat)
}

Kb_b.inv <- 1/Kb_b(n,beta0,eta0,theta0,x0)

Kb_b_b <- function(n, beta, eta0, theta0, x) {
csi <- beta*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*x*dnorm(csi)
a <- V1/qx
b <- V1/(1-qx)
mat <- (a+b)^3*qx-b^3-3*(a+b)^2*b*qx+3*(a+b)*b^2*qx
(1/n)*sum(mat)
}

Kb_bb <- function (n, beta, eta0, theta0, x) {
csi <- beta*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*x*dnorm(csi)
a <- V1/qx
b <- V1/(1-qx)
V2 <- (1-eta0-theta0)*x^2*dnorm(csi)*csi
c <- (V2*qx-V1^2)/qx^2
d <- (V2*(1-qx)+V1^2)/(1-qx)^2
mat <- (a+b)*(c+d)*qx-(a+b)*d*qx-(c+d)*b*qx+b*d
(1/n)*sum(mat)
}

Ab <- function(n,beta0,eta0,theta0,x0,y0,Kb_b.inv) {
-Ubb(y0,beta0,eta0,theta0,x0)/(2*n)*Kb_b.inv^2*
```

```
(Kb_b_b(n,beta0,eta0,theta0,x0)+Kb_bb(n,beta0,eta0,theta0,x0))
}

library(nleqslv)

# Con la funzione di punteggio modificata (Firth):

U.star <- function(beta0, eta0, theta0, x0, y0) {
n = length(y0)

Kb_b.inv <- 1/Kb_b(n,beta0,eta0,theta0,x0)

y = numeric(1)

ub = Ub(y0, beta0, eta0, theta0, x0)

ab = Ab(n,beta0,eta0,theta0,x0,y0,Kb_b.inv)

y = ub + ab

y
}

nleqslv(0,U.star,eta0=eta0,theta0=theta0,x0=x0,
y0=y0,control=list(btol=.01))

library(numDeriv)

stat1par.probit<-function(R,n,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999)) {
set.seed(123)
x0 <- rnorm(n)
std.error.beta<-wald<-rep(NA,R)
mat.est <- matrix(0,1,R)
for(i in 1:R) {
csi0 <- beta0*x0
pi0probit <- pnorm(csi0)
q0 <- 1-theta0-(1-eta0-theta0)*pi0probit
```

```
y0 <- rbinom(n,1,q0)
mat.est[,i] = nleqslv(0, U.star, eta0=eta0,
theta0=theta0, x0=x0, y0=y0, control=list(btol=.01))$x
std.error.beta[i] <- sqrt(-1/grad(U.star,mat.est[1,i],
eta0=eta0,theta0=theta0,y0=y0,x0=x0))
wald[i] <- (mat.est[1,i] - beta0)^2/std.error.beta[i]^2
}
wald.coverage <- sapply(conf.level, function(x0) mean(wald
[which(!is.na(wald))] < qchisq(x0, 1)))
list(Bias=apply(mat.est,1,mean)-beta0, Mat.est=mat.est,
Wald.coverage=wald.coverage,
std.error=apply(mat.est,1,function(x) sqrt(var(x))))
}

prova1par100probit <- stat1par.probit(5000,100,beta0,eta0,
theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova1par200probit <- stat1par.probit(5000,200,beta0,eta0,
theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova1par500probit <- stat1par.probit(5000,500,beta0,eta0,
theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova1par1000probit <- stat1par.probit(5000,1000,beta0,eta0,
theta0,conf.level=c(0.90,0.95,0.99,0.999))

save(prova1par100probit,file="prova1par100probit.RData")
save(prova1par200probit,file="prova1par200probit.RData")
save(prova1par500probit,file="prova1par500probit.RData")
save(prova1par1000probit,file="prova1par1000probit.RData")

# Con la funzione di punteggio non modificata:

U.star.distorta <- function(beta0, eta0, theta0, x0, y0) {
n = length(y0)

y = numeric(1)

ub = Ub(y0, beta0, eta0, theta0, x0)

y = ub
```

```
y
}

nleqslv(0, U.star.distorta, eta0=eta0, theta0=theta0,
x0=x0, y0=y0, control=list(btol=.01))

stat1par.nomod.probit<-function(R,n,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999)) {
set.seed(123)
x0 <- rnorm(n)
std.error.beta<-wald<-rep(NA,R)
mat.est <- matrix(0,1,R)
for(i in 1:R) {
csi0 <- beta0*x0
pi0probit <- pnorm(csi0)
q0 <- 1-theta0-(1-eta0-theta0)*pi0probit
y0 <- rbinom(n,1,q0)
mat.est[,i] = nleqslv(0, U.star.distorta, eta0=eta0,
theta0=theta0, x0=x0, y0=y0, control=list(btol=.01))$x
std.error.beta[i] <- sqrt(-1/Ubb(y0,mat.est[1,i],eta0,
theta0,x0))
wald[i] <- (mat.est[1,i] - beta0)^2/std.error.beta[i]^2
}
wald.coverage <- sapply(conf.level, function(x0) mean(wald
[which(!is.na(wald))] < qchisq(x0, 1)))
list(Bias=apply(mat.est,1,mean)-beta0, Mat.est=mat.est,
Wald.coverage=wald.coverage,
std.error=apply(mat.est,1,function(x) sqrt(var(x))))
}

prova2.1par100probit <- stat1par.nomod.probit(5000,100,beta0,
eta0,theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova2.1par200probit <- stat1par.nomod.probit(5000,200,beta0,
eta0,theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova2.1par500probit <- stat1par.nomod.probit(5000,500,beta0,
eta0,theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova2.1par1000probit <- stat1par.nomod.probit(5000,1000,beta0,
```

```
eta0,theta0,conf.level=c(0.90,0.95,0.99,0.999))

save(prova2.1par100probit,file="prova2.1par100probit.RData")
save(prova2.1par200probit,file="prova2.1par200probit.RData")
save(prova2.1par500probit,file="prova2.1par500probit.RData")
save(prova2.1par1000probit,file="prova2.1par1000probit.RData")


# PROBIT CON DUE PARAMETRI:

n <- 100
eta0 <- 0.90
theta0 <- 0.80
set.seed(123)
x0 <- rnorm(n)
beta1 <- 1
beta2 <- 1
beta0 <- c(beta1,beta2)

csi0 <- beta0[1]+beta0[2]*x0

pi0probit <- pnorm(csi0)
q0 <- 1-theta0-(1-eta0-theta0)*pi0probit

y0 <- rbinom(n,1,q0)

U1 <- function(y, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*dnorm(csi)
a1 <- V1/qx
b1 <- V1/(1-qx)
sum((a1+b1)*y)-sum(b1)
}

U2 <- function(y, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
```

```
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V2 <- -(1-eta0-theta0)*x*dnorm(csi)
a2 <- V2/qx
b2 <- V2/(1-qx)
sum((a2+b2)*y)-sum(b2)
}

U11 <- function(y, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*dnorm(csi)
V3 <- (1-eta0-theta0)*dnorm(csi)
c1 <- (V3*qx-V1^2)/qx^2
d1 <- (V3*(1-qx)+V1^2)/(1-qx)^2
sum((c1+d1)*y)-sum(d1)
}

U22 <- function(y, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V2 <- -(1-eta0-theta0)*x*dnorm(csi)
V4 <- (1-eta0-theta0)*x*dnorm(csi)
c2 <- (V4*qx-V2^2)/qx^2
d2 <- (V4*(1-qx)+V2^2)/(1-qx)^2
sum((c2+d2)*y)-sum(d2)
}

U12 <- function(y, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*dnorm(csi)
V2 <- -(1-eta0-theta0)*x*dnorm(csi)
V5 <- (1-eta0-theta0)*x*dnorm(csi)
c12 <- (V5*qx-V1*V2)/qx^2
```

```
d12 <- (V5*(1-qx)+V1*V2)/(1-qx)^2
sum((c12+d12)*y)-sum(d12)}

K1_1 <- function(n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*dnorm(csi)
a1 <- V1/qx
b1 <- V1/(1-qx)
mat <- (a1+b1)^2*qx+b1^2-2*(a1+b1)*b1*qx
(1/n)*sum(mat)
}

K2_2 <- function(n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V2 <- -(1-eta0-theta0)*x*dnorm(csi)
a2 <- V2/qx
b2 <- V2/(1-qx)
mat <- (a2+b2)^2*qx+b2^2-2*(a2+b2)*b2*qx
(1/n)*sum(mat)
}

K1_2 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*dnorm(csi)
a1 <- V1/qx
b1 <- V1/(1-qx)
V2 <- -(1-eta0-theta0)*x*dnorm(csi)
a2 <- V2/qx
b2 <- V2/(1-qx)
mat <- (a1+b1)*(a2+b2)*qx-(a1+b1)*b2*qx-(a2+b2)*b1*qx+b1*b2
(1/n)*sum(mat)
}
```

```
K <- matrix(c(K1_1(n,beta0,eta0,theta0,x0),
K1_2(n,beta0,eta0,theta0,x0),K1_2(n,beta0,eta0,theta0,x0),
K2_2(n,beta0,eta0,theta0,x0)),2,2)

Kinv <- solve(K)

K1_1_1 <- function(n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*dnorm(csi)
a1 <- V1/qx
b1 <- V1/(1-qx)
mat <- (a1+b1)^3*qx-b1^3-3*(a1+b1)^2*b1*qx+3*(a1+b1)*b1^2*qx
(1/n)*sum(mat)
}

K2_2_2 <- function(n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V2 <- -(1-eta0-theta0)*x*dnorm(csi)
a2 <- V2/qx
b2 <- V2/(1-qx)
mat <- (a2+b2)^3*qx-b2^3-3*(a2+b2)^2*b2*qx+3*(a2+b2)*b2^2*qx
(1/n)*sum(mat)
}

K1_11 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*dnorm(csi)
a1 <- V1/qx
b1 <- V1/(1-qx)
V3 <- (1-eta0-theta0)*dnorm(csi)
c1 <- (V3*qx-V1^2)/qx^2
```

```
d1 <- (V3*(1-qx)+V1^2)/(1-qx)^2
mat <- (a1+b1)*(c1+d1)*qx-(a1+b1)*d1*qx-(c1+d1)*b1*qx+b1*d1
(1/n)*sum(mat)
}

K1_1_2 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*dnorm(csi)
a1 <- V1/qx
b1 <- V1/(1-qx)
V2 <- -(1-eta0-theta0)*x*dnorm(csi)
a2 <- V2/qx
b2 <- V2/(1-qx)
mat <- (a1+b1)^2*(a2+b2)*qx-(a1+b1)^2*b2*qx+(a2+b2)*b1^2*qx-
b1^2*b2-2*(a1+b1)*b1*(a2+b2)*qx+2*(a1+b1)*b1*b2*qx
(1/n)*sum(mat)
}

K1_12 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*dnorm(csi)
a1 <- V1/qx
b1 <- V1/(1-qx)
V2 <- -(1-eta0-theta0)*x*dnorm(csi)
V5 <- (1-eta0-theta0)*x*dnorm(csi)
c12 <- (V5*qx-V1*V2)/qx^2
d12 <- (V5*(1-qx)+V1*V2)/(1-qx)^2
mat <- (a1+b1)*(c12+d12)*qx-(a1+b1)*d12*qx-(c12+d12)*b1*qx+b1*d12
(1/n)*sum(mat)
}

K1_2_2 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
```

```
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*dnorm(csi)
a1 <- V1/qx
b1 <- V1/(1-qx)
V2 <- -(1-eta0-theta0)*x*dnorm(csi)
a2 <- V2/qx
b2 <- V2/(1-qx)
mat <- (a1+b1)*(a2+b2)^2*qx-(a2+b2)^2*b1*qx+(a1+b1)*b2^2*qx-
b1*b2^2-2*(a2+b2)*b2*(a1+b1)*qx+2*(a2+b2)*b1*b2*qx
(1/n)*sum(mat)
}

K1_22 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V1 <- -(1-eta0-theta0)*dnorm(csi)
a1 <- V1/qx
b1 <- V1/(1-qx)
V2 <- -(1-eta0-theta0)*x*dnorm(csi)
V4 <- (1-eta0-theta0)*x*dnorm(csi)
c2 <- (V4*qx-V2^2)/qx^2
d2 <- (V4*(1-qx)+V2^2)/(1-qx)^2
mat <- (a1+b1)*(c2+d2)*qx-(a1+b1)*d2*qx-(c2+d2)*b1*qx+b1*d2
(1/n)*sum(mat)
}

K2_11 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V2 <- -(1-eta0-theta0)*x*dnorm(csi)
a2 <- V2/qx
b2 <- V2/(1-qx)
V1 <- -(1-eta0-theta0)*dnorm(csi)
V3 <- (1-eta0-theta0)*dnorm(csi)
c1 <- (V3*qx-V1^2)/qx^2
d1 <- (V3*(1-qx)+V1^2)/(1-qx)^2
```

```
mat <- (a2+b2)*(c1+d1)*qx-(a2+b2)*d1*qx-(c1+d1)*b2*qx+b2*d1
(1/n)*sum(mat)
}

K2_12 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V2 <- -(1-eta0-theta0)*x*dnorm(csi)
a2 <- V2/qx
b2 <- V2/(1-qx)
V1 <- -(1-eta0-theta0)*dnorm(csi)
V5 <- (1-eta0-theta0)*x*dnorm(csi)
c12 <- (V5*qx-V1*V2)/qx^2
d12 <- (V5*(1-qx)+V1*V2)/(1-qx)^2
mat <- (a2+b2)*(c12+d12)*qx-(a2+b2)*d12*qx-(c12+d12)*b2*qx+b2*d12
(1/n)*sum(mat)
}

K2_22 <- function (n, beta, eta0, theta0, x) {
csi <- beta[1]+beta[2]*x
pi_probit <- pnorm(csi)
qx <- 1-theta0-(1-eta0-theta0)*pi_probit
V2 <- -(1-eta0-theta0)*x*dnorm(csi)
a2 <- V2/qx
b2 <- V2/(1-qx)
V4 <- (1-eta0-theta0)*x*dnorm(csi)
c2 <- (V4*qx-V2^2)/qx^2
d2 <- (V4*(1-qx)+V2^2)/(1-qx)^2
mat <- (a2+b2)*(c2+d2)*qx-(a2+b2)*d2*qx-(c2+d2)*b2*qx+b2*d2
(1/n)*sum(mat)
}

A1 <- function(n,beta0,eta0,theta0,x0,y0,Kinv) {
-U11(y0,beta0,eta0,theta0,x0)/(2*n)*(Kinv[1,1]*Kinv[1,1]*
(K1_1_1(n,beta0,eta0,theta0,x0)+K1_11(n,beta0,eta0,theta0,x0))+
Kinv[1,1]*Kinv[1,2]*(K1_1_2(n,beta0,eta0,theta0,x0)+
K1_12(n,beta0,eta0,theta0,x0))+Kinv[1,1]*Kinv[1,2]*
```

```
(K1_1_2(n,beta0,eta0,theta0,x0)+K1_12(n,beta0,eta0,theta0,x0))+
Kinv[1,1]*Kinv[2,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K1_22(n,beta0,eta0,theta0,x0))+Kinv[1,2]*Kinv[1,1]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K2_11(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[1,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K2_12(n,beta0,eta0,theta0,x0))+Kinv[1,2]*Kinv[1,2]*
(K1_2_2(n,beta0,eta0,theta0,x0)+K2_12(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[2,2]*(K2_2_2(n,beta0,eta0,theta0,x0)+
K2_22(n,beta0,eta0,theta0,x0)))-
U12(y0,beta0,eta0,theta0,x0)/(2*n)*(Kinv[1,2]*Kinv[1,1]*
(K1_1_1(n,beta0,eta0,theta0,x0)+K1_11(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[1,2]*(K1_1_2(n,beta0,eta0,theta0,x0)+
K1_12(n,beta0,eta0,theta0,x0))+Kinv[1,2]*Kinv[1,2]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K1_12(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[2,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K1_22(n,beta0,eta0,theta0,x0))+Kinv[2,2]*Kinv[1,1]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K2_11(n,beta0,eta0,theta0,x0))+
Kinv[2,2]*Kinv[1,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K2_12(n,beta0,eta0,theta0,x0))+Kinv[2,2]*Kinv[1,2]*
(K1_2_2(n,beta0,eta0,theta0,x0)+K2_12(n,beta0,eta0,theta0,x0))+
Kinv[2,2]*Kinv[2,2]*(K2_2_2(n,beta0,eta0,theta0,x0)+
K2_22(n,beta0,eta0,theta0,x0)))
}

A2 <- function(n,beta0,eta0,theta0,x0,y0,Kinv) {
-U12(y0,beta0,eta0,theta0,x0)/(2*n)*(Kinv[1,1]*Kinv[1,1]*
(K1_1_1(n,beta0,eta0,theta0,x0)+K1_11(n,beta0,eta0,theta0,x0))+
Kinv[1,1]*Kinv[1,2]*(K1_1_2(n,beta0,eta0,theta0,x0)+
K1_12(n,beta0,eta0,theta0,x0))+Kinv[1,1]*Kinv[1,2]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K1_12(n,beta0,eta0,theta0,x0))+
Kinv[1,1]*Kinv[2,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K1_22(n,beta0,eta0,theta0,x0))+Kinv[1,2]*Kinv[1,1]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K2_11(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[1,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K2_12(n,beta0,eta0,theta0,x0))+Kinv[1,2]*Kinv[1,2]*
(K1_2_2(n,beta0,eta0,theta0,x0)+K2_12(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[2,2]*(K2_2_2(n,beta0,eta0,theta0,x0)+
K2_22(n,beta0,eta0,theta0,x0)))-
```

```
U22(y0,beta0,eta0,theta0,x0)/(2*n)*(Kinv[1,2]*Kinv[1,1]*
(K1_1_1(n,beta0,eta0,theta0,x0)+K1_11(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[1,2]*(K1_1_2(n,beta0,eta0,theta0,x0)+
K1_12(n,beta0,eta0,theta0,x0))+Kinv[1,2]*Kinv[1,2]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K1_12(n,beta0,eta0,theta0,x0))+
Kinv[1,2]*Kinv[2,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K1_22(n,beta0,eta0,theta0,x0))+Kinv[2,2]*Kinv[1,1]*
(K1_1_2(n,beta0,eta0,theta0,x0)+K2_11(n,beta0,eta0,theta0,x0))+
Kinv[2,2]*Kinv[1,2]*(K1_2_2(n,beta0,eta0,theta0,x0)+
K2_12(n,beta0,eta0,theta0,x0))+Kinv[2,2]*Kinv[1,2]*
(K1_2_2(n,beta0,eta0,theta0,x0)+K2_12(n,beta0,eta0,theta0,x0))+
Kinv[2,2]*Kinv[2,2]*(K2_2_2(n,beta0,eta0,theta0,x0)+
K2_22(n,beta0,eta0,theta0,x0)))
}

library(nleqslv)

# Con la funzione di punteggio modificata (Firth):

u.star.beta <- function(beta0, eta0, theta0, x0, y0) {
n = length(y0)

K <- matrix(c(K1_1(n,beta0,eta0,theta0,x0),
K1_2(n,beta0,eta0,theta0,x0),K1_2(n,beta0,eta0,theta0,x0),
K2_2(n,beta0,eta0,theta0,x0)),2,2)
Kinv <- solve(K)

y = numeric(2)

u1 = U1(y0, beta0, eta0, theta0, x0)
u2 = U2(y0, beta0, eta0, theta0, x0)

a1 = A1(n,beta0,eta0,theta0,x0,y0,Kinv)
a2 = A2(n,beta0,eta0,theta0,x0,y0,Kinv)

y[1] = u1 + a1
y[2] = u2 + a2
```

```
y
}

nleqslv(c(0,0), u.star.beta, eta0=eta0, theta0=theta0,
x0=x0, y0=y0, control=list(btol=.01))

library(numDeriv)

# Matrice di informazione osservata inversa:

funct.genD <- function(beta0, eta0, theta0, x0, y0) {
-u.star.beta(c(beta0[1],beta0[2]), eta0, theta0, x0, y0)
}

info.oss.firth.inv <- function(beta0, eta0, theta0, x0, y0){
ris<- genD(funct.genD,c(1,1),eta0=eta0,theta0=theta0,
x0=x0,y0=y0)$D[,1:2]
solve(ris)
}

statistiche.probit<-function(R,n,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999)) {
set.seed(123)
x0 <- rnorm(n)
mat.est <- matrix(0,2,R)
std.error.beta1<-std.error.beta2<-wald1<-wald2<-rep(NA,R)
for(i in 1:R) {
csi0 <- beta0[1]+beta0[2]*x0
pi0probit <- pnorm(csi0)
q0 <- 1-theta0-(1-eta0-theta0)*pi0probit
y0 <- rbinom(n,1,q0)
info <- sqrt(diag(info.oss.firth.inv(beta0,eta0,theta0,
x0,y0)))
std.error.beta1[i] <- info[1]
std.error.beta2[i] <- info[2]
mat.est[,i] = nleqslv(c(0,0), u.star.beta, eta0=eta0,
theta0=theta0, x0=x0, y0=y0, control=list(btol=.01))$x
wald1[i] <- (mat.est[1,i] - beta0[1])^2/std.error.beta1[i]^2
```

```
wald2[i] <- (mat.est[2,i] - beta0[2])^2/std.error.beta2[i]^2
}
wald.coverage1 <- sapply(conf.level, function(x0) mean(wald1
[which(!is.na(wald1))] < qchisq(x0, 1)))
wald.coverage2 <- sapply(conf.level, function(x0) mean(wald2
[which(!is.na(wald2))] < qchisq(x0, 1)))
list(Bias=apply(mat.est,1,mean)-beta0, Mat.est=mat.est,
Wald.coverage1=wald.coverage1,
Wald.coverage2=wald.coverage2,
std.error=apply(mat.est,1,function(x) sqrt(var(x))))
}

prova100probit <- statistiche.probit(5000,100,beta0,eta0,
theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova200probit <- statistiche.probit(5000,200,beta0,eta0,
theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova500probit <- statistiche.probit(5000,500,beta0,eta0,
theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova1000probit <- statistiche.probit(5000,1000,beta0,eta0,
theta0,conf.level=c(0.90,0.95,0.99,0.999))

save(prova100probit,file="prova100probit.RData")
save(prova200probit,file="prova200probit.RData")
save(prova500probit,file="prova500probit.RData")
save(prova1000probit,file="prova1000probit.RData")

# Con la funzione di punteggio non modificata:

u.star.beta.distorta<-function(beta0,eta0,theta0,x0,y0) {
n = length(y0)

y = numeric(2)

u1 = U1(y0, beta0, eta0, theta0, x0)
u2 = U2(y0, beta0, eta0, theta0, x0)

y[1] = u1
y[2] = u2
```

```
y
}

nleqslv(c(0,0), u.star.beta.distorta, eta0=eta0, theta0=theta0,
x0=x0, y0=y0, control=list(btol=.01))

# Matrice di informazione osservata inversa:

info.oss.inv <- function(beta0, eta0, theta0, x0, y0) {
n = length(y0)
solve(-matrix(c(U11(y0,beta0,eta0,theta0,x0),U12(y0,beta0,eta0,
theta0,x0),U12(y0,beta0,eta0,theta0,x0),
U22(y0,beta0,eta0,theta0,x0)),2,2))
}

statistiche.nomod.probit<-function(R,n,beta0,eta0,theta0,
conf.level=c(0.90,0.95,0.99,0.999)) {
set.seed(123)
x0 <- rnorm(n)
mat.est <- matrix(0,2,R)
std.error.beta1<-std.error.beta2<-wald1<-wald2<-rep(NA,R)
for(i in 1:R) {
csi0 <- beta0[1]+beta0[2]*x0
pi0probit <- pnorm(csi0)
q0 <- 1-theta0-(1-eta0-theta0)*pi0probit
y0 <- rbinom(n,1,q0)
info <- sqrt(diag(info.oss.inv(beta0,eta0,theta0,x0,y0)))
std.error.beta1[i] <- info[1]
std.error.beta2[i] <- info[2]
mat.est[,i] = nleqslv(c(0,0),u.star.beta.distorta,eta0=eta0,
theta0=theta0, x0=x0, y0=y0, control=list(btol=.01))$x
wald1[i] <- (mat.est[1,i] - beta0[1])^2/std.error.beta1[i]^2
wald2[i] <- (mat.est[2,i] - beta0[2])^2/std.error.beta2[i]^2
}
wald.coverage1 <- sapply(conf.level, function(x0) mean(wald1
[which(!is.na(wald1))] < qchisq(x0, 1)))
wald.coverage2 <- sapply(conf.level, function(x0) mean(wald2
```

```
[which(!is.na(wald2))] < qchisq(x0, 1)))
list(Bias=apply(mat.est,1,mean)-beta0, Mat.est=mat.est,
Wald.coverage1=wald.coverage1,
Wald.coverage2=wald.coverage2,
std.error=apply(mat.est,1,function(x) sqrt(var(x))))
}

prova2_100probit <- statistiche.nomod.probit(5000,100,beta0,
eta0,theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova2_200probit <- statistiche.nomod.probit(5000,200,beta0,
eta0,theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova2_500probit <- statistiche.nomod.probit(5000,500,beta0,
eta0,theta0,conf.level=c(0.90,0.95,0.99,0.999))
prova2_1000probit <- statistiche.nomod.probit(5000,1000,beta0,
eta0,theta0,conf.level=c(0.90,0.95,0.99,0.999))

save(prova2_100probit,file="prova2_100probit.RData")
save(prova2_200probit,file="prova2_200probit.RData")
save(prova2_500probit,file="prova2_500probit.RData")
save(prova2_1000probit,file="prova2_1000probit.RData")


# ELIMINO I CAMPIONI PROBLEMATICI DAL PROBIT:

# CON UN SOLO PARAMETRO:

load("prova2.1par100probit.RData")
ls()
names(prova2.1par100probit)
w1par = prova2.1par100probit$Mat.est
dim(w1par)

vect <- vector()
for(i in 1:5000) {
if(any(abs(w1par[,i])>10)) {
vect <- c(i,vect)
cat("i=",i,"w1par=",w1par[,i],"\n")
}
```

```
}

i= 1769 w1par= -92.86111
i= 1821 w1par= 862.0857
i= 2215 w1par= -922.5225
i= 3542 w1par= 19.36008
i= 4299 w1par= 54.02464

# Indici (campioni) problematici
vect

vect <- c(4299,3542,2215,1821,1769)

length(vect)

w1parbis <- w1par[,-vect]
length(w1parbis)

NuovaDist1par100 <- mean(w1parbis)-1
NuovaDist1par100

load("prova1par100probit.RData")

s1par <- prova1par100probit$Mat.est
s1par[,vect]

s1parbis <- s1par[,-vect]

Firth1par100 <- mean(s1parbis)-1
Firth1par100


# CON DUE PARAMETRI:

load("prova2_100probit.RData")
ls()
names(prova2_100probit)
w <- prova2_100probit$Mat.est
```

```
dim(w)

vect <- vector()
for(i in 1:5000) {
if(any(abs(w[,i])>20)) {
vect <- c(i,vect)
cat("i=",i,"w=",w[,i],"\n")
}
}

vect <- c(4976,4930,4813,4809,4703,4647,4606,4553,
4546,4545,4424,4181,4145,4044,4030,4027,3940,3803,
3745,3730,3696,3660,3494,3463,3452,3174,3118,3108,
3095,2929,2873,2786,2784,2773,2746,2729,2508,2450,
2437,2425,2405,2394,2388,2379,2302,2267,2010,1984,
1922,1734,1488,1446,1362,1077,1052,894,886,869,846,
562,498,443,343,285,202,94,41,26,18)

length(vect)

wbis <- w[,-vect]
dim(wbis)

prova2_100probit_bis <- apply(wbis,1,mean)-c(1,1)
prova2_100probit_bis

load("prova100probit.RData")
ls()
names(prova100probit)
s <- prova100probit$Mat.est
dim(s)

s[,vect]

sbis <- s[,-vect]
Firth100 <- apply(sbis,1,mean)-c(1,1)
Firth100
```

Riferimenti bibliografici

Azzalini, A. (2001). *Inferenza Statistica - Una presentazione basata sul concetto di verosimiglianza*. 2nd ed., Milano: Springer Verlag.

Barron, B.A. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics*, **33**, 414-417.

Brenner, H. & Gefeller, O. (1993). Use of positive predictive value to correct for disease misclassification in epidemiologic studies. *Am. J. Epidem.*, **138**, 1007-1015.

Carroll, R.J., Ruppert, D., Stefanski, L.A. & Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Copeland, K.T., Checkoway, H., McMichael, A.J. & Holbrook, R.H. (1977). Bias due to misclassification in the estimation of relative risk. *Am. J. Epidem.*, **105**, 488-495.

Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, **7**, 1-26.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27-38.

Flegal, K.M., Brownie, C. & Haas, J.D. (1986). The effects of exposure misclassification on estimates of relative risk. *Am. J. Epidem.*, **123**, 736-751.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*. New york: 2nd ed. John Wiley.

Franco, E.L. (1992). Measurement errors in epidemiological studies of human papillomavirus and cervical cancer. In *The Epidemiology of Cervical Cancer and Human Papillomavirus*, IARC Scientific Publication 119, Ed.N. Munoz et al., pp. 181-197. Lyon: International Agency for Research on Cancer.

Green, M.S. (1983). Use of predicative value to adjust relative risk estimates biased by misclassification of outcome status. *Am. J. Epidem.*, **117**, 98-105.

Greenland, S. (1988). Variance estimation of epidemiologic effect estimates under misclassification. *Statist. Med.*, **7**, 745-757.

Küchenhoff, H. (2007). Materiale per il corso *Misclassification and measurement error in regression models* [2,3 ottobre 2007]. Scuola di Dottorato in Scienze Statistiche, Università di Padova.

Magder, L.S. & Hughes, J.P. (1997). Logistic regression when the outcome is measured with uncertainty. *Am. J. Epidem.*, **146**, 195-203.

McCullagh, P. (1987). *Tensor Methods in Statistics*. London: Chapman and Hall.

McInturff, P., O Johnson, W., Cowling, D. & A Gardner, I. (2004). Modelling risk when binary outcomes are subject to error. *Statist. Med.*, **23**, 1095-1109.

Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models. *J. Roy. Statist. Soc.* **A, 135**, 370-384.

Neuhaus, J.M. (1999). Bias and Efficiency Loss Due to Misclassified Responses in Binary Regression. *Biometrika*, **86**, 843-855.

Pace, L. e Salvan, A. (2001). *Introduzione alla Statistica II - Inferenza, verosimiglianza, modelli*. Padova: Cedam.

Pace, L. e Salvan, A. (1996). *Teoria della Statistica: Metodi, Modelli, Approssimazioni Asintotiche*. Padova: Cedam.

Piccolo, D. (2006). *Statistica per le decisioni - La conoscenza umana sostenuta dall'evidenza empirica*. Bologna: Il Mulino.

Quenouille, M.H. (1949). Approximate tests of correlation in time-series. *J. R. Statist. Soc.*, **B, 11**, 68-84.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika*, **43**, 353-360.

Ramsey, F.P. (1931). *The Foundations of Mathematics and Other Logical Essays*. Kegan Paul, Trench, Trubner & Co. Ltd, London.

Sartori, N. (2013). Materiale didattico per il corso *Statistica Computazionale (progredito)*. Dipartimento di Scienze Statistiche, Università di Padova.

Scholz, F.W. (2007). Materiale per il corso *The bootstrap small sample properties* [25 giugno 2007]. University of Washington.

Weinberg, C.R., Umbach, D.M. & Greenland, S. (1994). When will nondifferential misclassification of an exposure preserve the direction of a trend? *Am. J. Epidem.*, **140**, 565-571.